

Time to unite Islam and science

Science in Muslim countries is weak. The reasons for this deserve attention, as do the consequences for these nations' economic health. A meeting last week provided a start in this direction.

"I am a Muslim, and we Muslims don't usually speak out on such matters. But I'm going to, as I have a responsibility towards our children, society, education and the advancement of scientific thinking." The seemingly taboo question that Çigdem Kagıtcıbası, a psychologist and member of the Turkish Academy of Sciences, was nervously addressing was whether a lack of democracy and freedom of expression in the Islamic world might be one reason why there is no substantial culture of science and technology in these countries.

Addressing a satellite workshop on 'science, religion and values' at a meeting of top science officials from Islamic and Western countries last week in Trieste, Italy (see page 101), Kagıtcıbası argued the need to tackle head-on this question: "Is 'Muslim culture' helping the advancement of our society or not?"

Provocatively, and perhaps undiplomatically given her secular Turkish background, she questioned, for example, Saudi Arabia's sponsorship of Wahhabism, a particularly rigid and austere form of Islam, at mosques and madrassas, or Islamic religious schools, worldwide. "Why are they opening up religious schools everywhere instead of ordinary schools devoted to basic education?" she asked.

Mohamed Falougi, deputy director general of the Islamic Educational, Scientific and Cultural Organization, protested that such discussion was "inappropriate" and that Kagıtcıbası had no right to say such things: "Islamic countries can teach any way they want, just as Judaism or Christianity do." It is true that explanations for the scientific ills in Islamic countries are complex. But the time has come to scrutinize Islam's relationship, inhibitory or otherwise, with science.

Neglecting the impact of Islam on science would not only be blind, but a disservice to Muslim peoples, who, if they are to become prosperous, need to shift from their flagging natural resource- and agriculture-based economies to knowledge-based ones. The Islamic world's science spending is just 0.2% of gross national product, its population of scientists is meagre, and its legal framework for innovation largely non-existent. Some countries, such as Pakistan, have recently shifted towards a culture of technology and scientific knowledge, but this is an exception. Its visionary, the higher-education minister, Atta-ur-Rahman, is an outspoken proponent of the need for a revolution in the Islamic world's commitment to science and education.

The Koran encourages the pursuit of science, and in its heyday the Islamic world was a cradle of science for six centuries. But many scientists, including Rahman, think that aspects of contemporary Muslim cultures are hindering the emergence of scientific cultures through narrow-mindedness, lack of basic freedoms, and pressure on individuals to conform, and not to challenge or think critically and creatively. Greater dialogue is needed among Muslim scholars and scientists about how to encourage science to flourish in the Islamic world.

Poverty, unemployment, the Middle East crisis and a perception of injustice with respect to the actions of the West are creating a cycle of violence and hatred, which often finds its expression in radical Islam. Science and education, the freedom to think critically, and contact with Western scientists can help to educate both sides, leading to a reform of Islam that brings to the fore the tolerance and scholarship that was for centuries the mark of the religion's practice. ■

A little protectionism goes a long way

China's stalling on introducing transgenic crops may frustrate outsiders, but helps it nurture its own biotechnology industry.

It is highly unusual for a developing country to assume a leading position in an important field of scientific endeavour. But China says that it intends to do just that in agricultural biotechnology — and all the signs are that its plan should be taken seriously.

China has already devoted a considerable research effort to plant science, including the development of transgenic crops (see page 111). Its public programme is making progress on many fronts, from sequencing the rice genome to producing hundreds of promising crop varieties. The programme is characterized not just by the nation's customary determination, but also by talent, imagination and a strong outward orientation that engages Chinese scientists in collaboration with colleagues overseas.

Amid all this activity, Chinese regulators have recently been reluctant to approve the commercial planting of transgenic food crops (genetically modified cotton is already well established). The government claims to be worried about whether Europe will accept exports of transgenic food, and about public reaction at home. Neither argument is convincing: its agricultural trade with Europe is small, and the public reaction is firmly under the state's control.

What China really wants is some breathing space to enable its own transgenic technology to catch up with that abroad. There's nothing

inherently wrong with this: unless the United States wants to fight a war to open up China's markets, as Britain once did to force it to accept opium imports, it needs to recognize China's strategic need to develop its own agricultural biotechnology. Advocates of US agricultural biotechnology, such as Senator Chuck Grassley (Republican, Iowa), probably want the United States to take China to the World Trade Organization for protecting its markets in this way. But this heavy-handed approach will do the technology no favours.

Like all developing countries, China needs time to assess and develop agricultural biotechnology on its own terms. The stranglehold on patents for methods and genes that Western corporations and universities have obtained in this arena is already making life hard for agricultural scientists across the developing world. At least China has the political and financial clout to overcome these obstacles and develop the technology in the interests of its own farmers.

An old Chinese proverb says that if you sit on the banks of the river long enough, the bodies of your enemies will float down past you. It is not clear if the author had President Bush, Senator Grassley or the board of Monsanto in mind. But China's patience will, in the long run, be the best way to assure the fruitful introduction of agricultural biotechnology into the world's most populous nation. ■





The bottom line
Michigan's biotech dream falls foul of funding cuts
p102



Poles apart
Report sparks row over effects of oil drilling in the Arctic
p103



Cash crisis
Max Planck Society pulls plug on chemical units
p105



Secret weapon
US Army seeks partners for biotech research
p106

Academies wrestle with issue of Islam's flagging science base

Declan Butler, Trieste

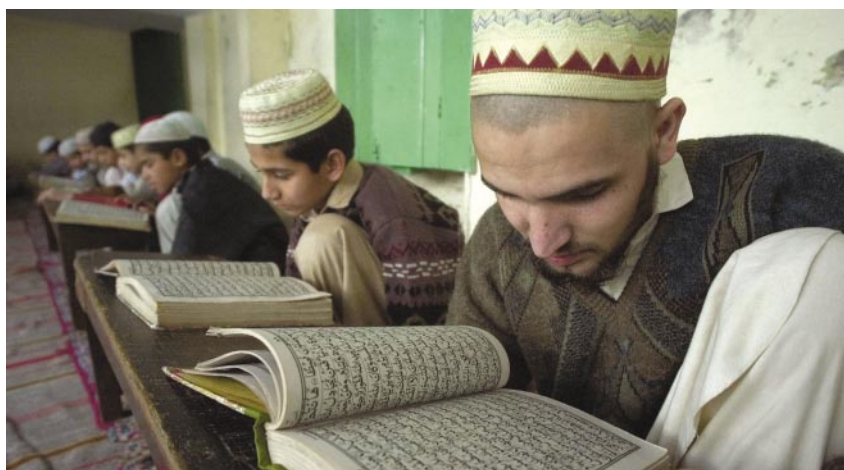
With the world on the brink of a conflict that is likely to reverberate for years to come, scientific leaders gathered in Trieste, Italy, last week to discuss why Islamic countries have neglected science and technology, and what could be done to end this neglect.

Stretching from Indonesia to Morocco, and from Uganda to Kazakhstan, Muslim countries are home to 1.3 billion people and three-quarters of the world's fuel reserves. But their combined gross national product (GNP) is less than half that of Germany, illiteracy levels are among the world's highest, and science spending, at 0.2% of GNP, is well below the global average.

The meeting of research ministers and heads of scientific academies was organized by the InterAcademy Panel on International Issues, as part of an effort to strengthen the resources and roles of scientific academies worldwide in promoting research and providing independent advice to governments.

The delegates argued that political leaders in Islamic nations often fail to appreciate the importance of scientific research to their countries' development. Most speakers blamed public spending skewed towards the military, low educational standards, and a lack of public interest in science for this pattern.

But there was less agreement on whether



The Koran states that Muslims should seek knowledge of nature, so why is science being neglected?

the region's stunted scientific development can be linked to the practice of Islam. Many speakers pointed out that the religion itself is pro-science, with the Koran stating the duty of Muslims to seek knowledge of nature. Indeed, they added that the Islamic world enjoyed six centuries of scientific progress while Europe was floundering in the Dark Ages.

But some delegates argued that aspects of Islam — particularly fundamentalism — are incompatible with modernity and science.

They even proposed that scientists should help to 'reform' Islam. Others said that better science education might discourage people from what many scientists regard as the irrational aspects of religious belief.

"The problem is coming from dogma, from narrow interpretations of the Koran," said Atta-ur-Rahman, president of the Pakistan Academy of Sciences and minister for higher education. Rahman, who is director of the Husein Ebrahim Jamal Research

Bush seeks to beef up US nuclear-weapons arsenal

Geoff Brumfiel, Washington

The Bush administration is mounting a fresh push to add to the United States' nuclear-weapons arsenal.

In a report to go to Congress this week, the Air Force is expected to make the case for developing a "robust nuclear earth penetrator" that can destroy deeply buried targets. The earth penetrator, proposed last year by Bush-administration officials (see *Nature* 415, 945; 2002), would be a hardened version of an existing warhead.

The administration has also taken steps to develop new, low-yield nuclear weapons,

which it says could be used to destroy chemical and biological weapon facilities. The president's 2004 budget asks for the repeal of a 1993 law that prohibits the development of weapons with yields of less than 5 kilotonnes as they would lower the threshold for using nuclear weapons.

In a hearing last week, Congressman Mac Thornberry (Republican, Texas) said the new weapons would strengthen the United States' nuclear deterrent. "I'm concerned about how we keep our deterrent credible," he said.

But scientists familiar with nuclear-weapon design questioned the usefulness of

the proposal and said that it would encourage global nuclear proliferation. "Penetrators can't penetrate very far," says Richard Garwin, a physicist who has advised the US government on nuclear-weapons policy. "And they will not destroy chemicals or even anthrax under likely conditions of storage."

Congresswoman Ellen Tauscher (Democrat, California) says she hopes to muster enough opposition to block the new weapons in Congress. "I think Democrats are getting increasingly fired up about this," she says. "And I think the American people should be fired up about it too."

Institute of Chemistry in Karachi, has championed plans to create a multibillion Pan Islamic R&D Fund by the 57-state Organization of the Islamic Conference (see *Nature* **416**, 120–122; 2002).

Perhaps the most contentious suggestion was that science is stifled by a lack of democracy in authoritarian Islamic countries. Adnan Badran, president of the Beirut-based Arab Academy of Sciences and former deputy director-general of the United Nations Educational, Scientific and Cultural Organization, is convinced that a lack of free expression and creative thinking explains the absence of a vibrant scientific culture. “Oppressive regimes are hindering the creativity of scientists in Islamic states, imprisoning the brains of the élite,” he said.

Whether religious culture helps or hinders the advancement of Islamic society is a “legitimate question”, argued Çigdem Kagitcibasi, a psychologist and member of the Turkish Academy of Sciences. She said it is important to assess the impact of religious dogma on education, and expressed concern about the growth of fundamentalist religious schools.

Such arguments drew fire from Mohamed Falougi, deputy director general of the Islamic Educational, Scientific and Cultural Organization, who said that they are political, and should not arise at a meeting of scientists. Saleh Al-Athel, president of the King Abdulaziz City for Science & Technology in Riyadh, Saudi Arabia, warned against the idea that science is synonymous with progress, whereas Islam is somehow ‘backward’.

Although the academic élite of Islamic countries is generally suspicious of the United States’ motives and commitment to democracy in the region, many scientists and intellectuals are now so impatient for greater freedom that they are willing to contemplate change imposed from outside, according to Badran.

The meeting provided a rare opportunity for Western scientific leaders, such as Bruce Alberts, president of the US National Academy of Sciences, to rub shoulders with their Islamic counterparts. Most said that the gathering marked a positive step towards greater cooperation between Islamic and Western scientists.

Rahman added that scientists are well placed to engage in dialogue between Islamic countries and the West. This was echoed by Abdel-Salam Majali, president of the Islamic Academy of Sciences and a former prime minister of Jordan. “Our scientific community needs to work hard at projecting the true image of our Islamic faith, and counter the gross distortions that have become part of Western mentality toward Islam since the events of 9/11,” he said. ■

Biotech project in turmoil as Michigan balances books

Jonathan Knight, San Francisco

A project to turn Michigan into a biotechnology powerhouse is being gutted to help keep the state out of debt.

The Life Sciences Corridor — established in 1999 with settlement money from a multi-state lawsuit against the tobacco industry — provides nearly \$50 million a year in research funding to Michigan’s major research universities and biotechnology start-ups (see *Nature* **400**, 391; 1999). In the long term, the project is intended to help lessen the state’s economic reliance on the motor industry.

But in February, Michigan’s governor, Jennifer Granholm, chopped the programme’s 2003 budget down to \$32.5 million as part of a package of cuts to address the state’s \$285-million shortfall. The deficit is projected to grow dramatically, and on 6 March Granholm proposed cutting all but \$20 million from the programme in 2004.

Supporters of the project were surprised by the governor’s decision. “In terms of creating jobs and sources of revenue, this was a rather creative proposal,” says Frank Press, a former president of the National Academy of Sciences and now a consultant with the Washington Advisory Group in Washington DC, which advised on early project planning. “To regress now is short-sighted.”

The Life Sciences Corridor was set up as a 20-year project to foster the sort of growth in commercial biotechnology that has surrounded universities in other states, notably California and Massachusetts.

Nearly half of the programme’s annual



Michigan State University is among the institutions set to lose biotechnology funding.

budget, administered by the Michigan Economic Development Corporation, has gone to fund basic research at public institutions such as Michigan State University, with the rest helping private start-up companies or public–private collaborations.

So far the fund has been involved in some 50 Michigan start-ups. The declining economy has made the project’s job even more important, says its managing director, Raili Kerppola. “We have been playing the critical role of providing seed capital to new companies while private-sector investment has pretty much shut down,” he says.

Grants from the Life Sciences Corridor are given in lump sums, so previous awards will remain untouched, Kerppola says. But far fewer awards will be made this year.

Granholm said that cuts were a “tough choice” forced by Michigan voters’ resistance to tax increases. ■

Auction of DNA archive cancelled

Rex Dalton, San Diego

A historic archive of molecular-biology papers will not be broken up and sold, as planned, at auction next month. Jeremy Norman, the California-based owner of the archive, said that he hopes instead to keep it intact and sell it to a library.

The auction house Christie’s announced on 4 March that it was cancelling the auction in New York of the Jeremy Norman Molecular Biology Archive. Including papers from Aaron Klug, Max Perutz, Rosalind Franklin, Francis Crick and James Watson, the archive was valued for Christie’s at between \$2.2 million and \$3.3 million.

“Christie’s is offering my collection *en bloc* to various institutional libraries in England and the United States,” Norman said in a statement. “Several institutions

have expressed interest and we are pursuing a private sale.”

“I am glad that Christie’s has withdrawn the auction,” says Klug, who sold his papers to the archive on the basis of an agreement that they would be kept together.

When the proposed sale, scheduled for 25 April, was disclosed last month, Klug and others expressed anger that the papers were to be split up (see *Nature* **421**, 564; 2003).

Al Seckel, a neuroscientist at the California Institute of Technology in Pasadena who purchased the papers for Norman, said that the auction was cancelled after his extensive documentation prohibiting breaking up the archive was brought to the attention of Christie’s. Neither Christie’s nor Norman would comment on why the sale was cancelled. ■

Senator rebuffs academy over Arctic oil

Tony Reichhardt, Washington

Thirty-five years of oil exploration and drilling on Alaska's northern coast has harmed the environment but had some positive effects on the human population, says a US National Academy of Sciences study.

According to the report, opening up the Arctic National Wildlife Refuge (ANWR) to more drilling would inflict further damage on fragile ecosystems. The finding follows a two-year study by a panel chaired by Gordon Orians, emeritus professor of zoology at the University of Washington.

The report was praised by opponents and some supporters of drilling. But Senator Ted Stevens, the pro-drilling Alaska Republican and appropriations committee chairman who commissioned the study in 1999, dismissed it as "another tool to question development" in the refuge. He impugned the neutrality of the panel, which included oil-industry consultants as well as wildlife ecologists, economists and social scientists.

But Stanley Senner, director of the National Audubon Society's Alaska office, said that the report could influence "the hearts and minds of the people still in the middle" of the ANWR debate.

The argument about drilling in the refuge is America's highest-profile environmental debate. Current law prohibits drilling or prospecting, but President George W. Bush and key Republicans want to open up part of the refuge to oil production. The debate should climax later this year, when Bush tries to push his plan through Congress.



Outlook unsettled: a caribou enjoys summer grazing, but oil exploration could threaten its habitat.

"There is no way to engage in these activities and have zero effect on the environment", Orians told a press briefing on 4 March. The 450-page report shows that most oil spills in Alaska have done little harm, and that modern oil-extraction technology is less harmful to the landscape than past methods. But three-dimensional seismic surveys, which have replaced earlier two-dimensional techniques, require more use of off-road trails, threatening the tundra's ecosystem.

Effects on caribou herds are hard to assess because of lack of data, uncertainty about where future drilling would take place, and the natural factors, such as insect activity, that also influence herds' behaviour. There has

been no large decline in the Central Arctic caribou herd, says the report, but the spread of industrial activity is likely to change this.

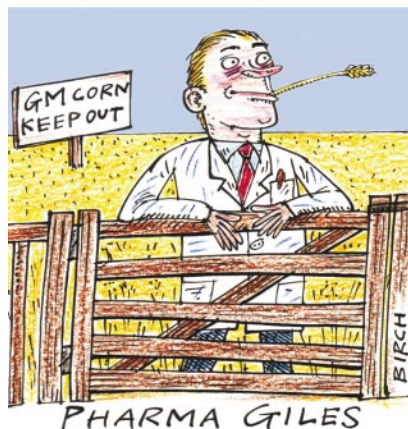
Tribal people have benefited from better health care and other services, the report adds. But the destruction of traditions may have contributed to social problems such as alcoholism. In some cases, human and ecological effects are intertwined: seismic exploration has driven bowhead whales to change their migration routes, forcing native hunters to take longer and riskier trips to catch them.

The panel recommends more research in the region, perhaps in study areas modelled on the National Science Foundation's Long-Term Ecological Research stations. ■

Tougher rules aim to prevent gene flow into crops

Hannah Hoag, Washington

US biotechnology companies will be allowed to grow genetically modified corn to produce pharmaceuticals — provided that it is grown at least a mile away from food crops and



is carefully monitored by government inspectors.

Tighter regulations for trials of 'pharma crops' were announced on 6 March by the US Department of Agriculture's Animal and Plant Health Inspection Service (APHIS). Last year, food corn and soy beans in Iowa and Nebraska had to be destroyed after possible contamination by corn that was being grown by ProdiGene of College Station, Texas, to produce a pig vaccine.

As well as doubling the required safety distance from food crops, the rules require dedicated machinery, extra training for planters, and up to five annual visits to each site by inspectors.

Some biotech companies have advocated excluding pharma crops from the corn belt in the American Midwest to prevent any possible public backlash. There are no plans for this, but the new regulations could have a similar effect, says Rebecca Bech, director

for biotechnology and regulatory services at APHIS. The safety zones should keep most trials out of the corn belt, she notes.

Biotech companies hope the regulations will help to build public confidence. "They are based on science and we support them strongly," says Debra Robertson, director of intellectual property at Epicyte, a pharma company in San Diego, which is developing corn to counteract the herpesvirus.

"Strong confinement conditions are only half of the battle," counters Gregory Jaffe, director of the biotechnology project at the Center for Science in the Public Interest in Washington DC. "You need to increase oversight and make the companies comply."

Both allies and critics are concerned that many of APHIS's 2,600 inspectors have been moved to the new Department of Homeland Security. "We've been assured that when the time comes the inspectors will be made available," says Bech. ■

Hydrogen cars fuel debate on basic research

Geoff Brumfiel, Washington

Experts are questioning the lack of university involvement in President Bush's initiative to produce a competitive hydrogen-fuelled car by 2020.

The programme was launched early last year and was augmented in January when Bush announced a \$1.7-billion, five-year research initiative to develop the technology and infrastructure to meet the project's goals.

When both houses of Congress held hearings last week to review the programme, no academic scientist was invited. Instead, committees in the House and the Senate asked technical managers of companies such as Shell and General Motors, together with environmental and industrial lobbyists, to testify.

But experts say that the programme — dubbed FreedomCAR (Cooperative Automotive Research) — needs a strong basic-research component if it is to come close to its ambitious goals. "Universities should have a much, much larger role in FreedomCAR," says Daniel Sperling, director of the Institute of Transportation Studies at the University of California, Davis.

FreedomCAR focuses on industrial research and development, as did the Partnership for a New Generation of Vehicles (PNGV) set up under Bill Clinton's administration, which it replaces (see *Nature* 415,



On the road: President Bush wants to see hydrogen-powered cars become a commercial reality.

248; 2002). But unlike the PNGV, which sought to increase the efficiency of existing technologies and components, the Bush programme will require an entirely new set of hydrogen-based technologies.

It aims to develop vehicles that will run on battery-like hydrogen fuel cells. Another component of the programme, the Hydrogen Fuel Initiative, would develop the infrastructure to transport and sell hydrogen fuel for the cars.

"The basic science underlying hydrogen fuel cells is very weak," says Sperling. Major

problems still dog the efficiency and reliability of the cells, and their integration into vehicle design. Over a dozen US universities have active fuel-cell research programmes.

"Universities need to be more involved," agrees Vernon Roan, who directs the fuel-cell lab at the University of Florida in Gainesville. His lab is currently testing a small bus that runs entirely on hydrogen fuel cells.

Similar basic issues surround the Hydrogen Fuel Initiative, says Malcolm Weiss, a transportation specialist at the Massachusetts Institute of Technology. Separating hydrogen from sources such as natural gas produces nearly as much greenhouse gas as petroleum fuels, he says, and hydrogen gas cannot be moved through conventional pipelines. That means that it may be necessary to produce hydrogen at the pump, perhaps through electrolysis of water. But the technologies to do this cheaply do not yet exist.

Nevertheless, the congressional committees were generally positive about the hydrogen-car programme — although some members expressed concern that its funding might come at the expense of other research into renewable energy.

Others, such as Senator Byron Dorgan (Democrat, North Dakota), said that the programme needed far more funding to be effective. "It needs to be an Apollo-type programme," Dorgan said, adding that he didn't want to be overly critical of the administration's rare foray into research on new energy sources. "When elephants fly," he said, "you ought not criticize them for being awkward."

"The impression I get from FreedomCAR is that they set objectives but there's no indication of how those objectives will be completed," says Weiss. Nevertheless, Sperling and Roan remain generally positive about the president's proposal. "From a long-term planning perspective it's an undoubtedly good programme," says Roan.

Russia pulls out of Antarctic station

Bryon MacWilliams, Moscow

The 13 crew members expected to arrive this week at Russia's Mirny Observatory on the Antarctic coast are likely to be tired and cold, and with good reason. When the Vostok research base at which they work ran out of fuel and supplies last month, the team was forced to take their Soviet-era trucks on a 1,400-kilometre journey across one of the coldest regions on Earth.

Vostok, which is near to the south geomagnetic pole, was abandoned for the winter on 28 February after bad weather prevented a convoy carrying fuel and supplies from reaching it. Valery Lukin of the Arctic and Antarctic Research Institute in St Petersburg, says that trips to the base have been disrupted since a supply ship attempting to reach Mirny became trapped in ice last year.

The crew's convoy was briefly stranded on 5 March when one of the trucks broke down, but it is expected to arrive at Mirny by 15 March. The same personnel are scheduled to reopen the base in early November, at the start of Antarctica's spring.

Lukin says that a proposed international project to drill into Lake Vostok, 4 km beneath the ice (see *Nature* 415, 828–830; 2002) will not be affected. Biologists suspect that the lake, which has been isolated for millions of years, could contain life.

The base has shut three times before in its 45-year history, although two of those closures were in the past decade. Lukin denies suggestions that a lack of government money is to blame. "This in no way means that Russia can't financially support its base," he says. "It is related to the climatic conditions."



Russia's Vostok station is closed for the winter.

Max Planck plans double blow to chemistry

Quirin Schiermeier, Munich

Two highly rated chemistry groups look set to become the first casualties of a budget freeze at Germany's Max Planck Society.

Nature has learned that plans to close the cosmochemistry and geochemistry departments at the Max Planck Institute for Chemistry in Mainz are expected to be approved by the society's ruling body in June — a move that some say would end Germany's involvement in top-quality science in these fields.

Up to 20 Max Planck research units are likely to shut down over the next few years. Peter Gruss, the society's president, said last November that such closures are unavoidable following the government's decision to freeze the society's budget at €1.2 billion (US\$1.3 billion) this year (see *Nature* **420**, 452; 2002).

The departments are being dismantled partly because their present scientific directors are due to retire. Geochemist Albrecht Hofmann will step down in 2005 and Günther Lugmair, head of cosmochemistry, will follow suit in 2007. The Max Planck Society traditionally builds departments around individual researchers, and always evaluates the future of a unit when its head leaves.

The Institute for Chemistry has proposed compensating for the closures by expanding its activities in atmospheric and climate research, with a special emphasis on remote satellite sensing. This new research structure has been agreed by a Max Planck committee, and is expected to be officially approved at a meeting of the society's senate in June.

But German researchers are worried by the plans. There are few other cosmochemistry and geochemistry groups in Germany, and the chemistry institute is widely acknowledged as one of Europe's premier centres for studies of the chemistry of the Earth's mantle and crust.

The cosmochemistry department also has an outstanding reputation in its own right. Lugmair and his group study the decay products of radionuclides found in meteorites to investigate planetary formation in the early Solar System. They operate one of the world's only two nanometre-resolution ion microprobes, devices that can detect tiny amounts of certain types of isotope. The €2-million facility was installed at the institute only two years ago. The department is also involved in planetary missions planned by NASA and the European Space Agency.

"The closures would basically end Germany's presence in cosmochemistry and geochemistry," says Alex Halliday, a geochemist at the Swiss Federal Institute of Technology in Zurich. Friedrich Seifert, a geophysicist at the Bavarian Research Institute for Experimental Geochemistry and Geophysics in Bayreuth, agrees. "It would be



Unplugged: installed just two years ago, this €2-million ion microprobe may soon be without a home.

a disaster for the Earth sciences, not only in Germany," he says.

Bernd Wirsing, spokesman for the Max Planck Society, says that scientific factors remain the main consideration when implementing cuts. But Lugmair is disappointed that the society is not even trying to find possible successors for him and Hofmann.

Before he joined the society in 1996, Lugmair was head of a cosmochemistry group at the Scripps Institution of Oceanography in San Diego, California. "If I had known how things would develop here in Germany, I would probably not have come," he says. ■

► www.mpg-mainz.mpg.de/mpg/english/index.html

Danish biotech centre faces axe

Alison Abbott, Munich

The Danish Biotechnology Instrument Center (DABIC), set up in 2000 to support research into genomics and proteomics, could face closure for want of funds to cover its running costs.

Based at the University of Southern Denmark in Odense, the centre was established with a three-year, €18-million (US\$20-million) grant. But it may now fall victim to a failed government policy.

In the late 1990s, the Danish government introduced about 50 highly specific research programmes, chosen by the national research councils, at a total cost of about €270 million. The government had hoped that universities would take over the running costs of successful centres created under the scheme. But the universities say they cannot afford to do so, after the government and its successor (elected in 2001) failed to increase their budgets even in line with inflation.

The centres see little hope in turning to the national research councils for help. Although the councils are enjoying a temporary increase in funding — thanks to a share of profits from the sale of

telecommunications licences — they cannot allocate the operational money needed to keep big centres going.

The DABIC is the largest of the centres. It involves 16 research groups across five universities, each of which bought expensive equipment and facilities. "Many of our groups will simply not be able to keep this equipment in service," says Peter Roepstorff, the centre's director. "Already we are losing people with expertise."

Jens Christian Djurhuus, head of Denmark's medical research council and director of the Institute for Experimental Clinical Research at the University of Aarhus, which hosts one of the DABIC's groups, says that his dual role makes him "ambivalent" about the situation. "The research councils were never particularly happy about having so many programmes with money earmarked for very specific projects, because there was not enough competition for each of them," he says. "But we were also unhappy when the former government lost interest in them, and we are worried about how the best ones can continue now." ■

Genetic-engineering projects falling by the wayside in Europe

Munich A litany of cancelled projects has been revealed by a European Commission survey of research into genetically modified plants and animals.

The survey, which will be published later this month, includes data from 168 public and private research units in 17 European countries. Some 39% of respondents reported cancelling projects on transgenic organisms in the past four years. The number of cancellations was higher in the private sector (61%) than among public research institutes and universities (23%). Over the same period, notifications for field trials of genetically modified plants dropped by around two-thirds.

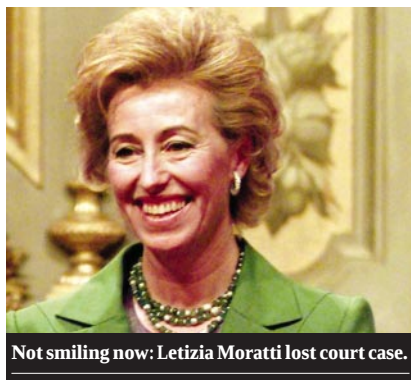
Respondents who cancelled projects cited consumer scepticism, unclear legislation and an uncertain market. The survey's organizer, Klaus Menrad of the Fraunhofer Institute for Systems and Innovation Research in Karlsruhe, Germany, says that such problems are forcing smaller European agribusinesses to look at alternative ways to produce new seeds, such as conventional plant breeding.

Italian science minister loses out over research reform

Rome Italian scientists had cause for relief last week, after a court rescued them from the threat of unwelcome reform.

A regional court in Rome ruled that research minister Letizia Moratti had acted beyond her power in publishing a decree in January that placed the CNR, Italy's main basic-research organization, in the hands of a commissioner (see *Nature* 421, 465; 2003). Moratti had said the decree was justified as senior CNR staff were unsympathetic to her plans to restructure the organization.

The case was brought by CNR president Lucio Bianco, whose position was to have been replaced by the commissioner, named as Adriano De Maio, rector of the private Luiss Guido Carli University in Rome. The court stated that the restructuring had not been



Not smiling now: Letizia Moratti lost court case.

made law and that the minister could not prejudice Bianco's willingness to implement it if required to do so — even though Bianco has publicly criticized the plans.

Laser quest foiled as officials sent packing

San Francisco The latest stage in a four-year spat over a \$20-million laser has ended with government officials being refused entry to Duke University in Durham, North Carolina.

The officials were rebuffed on 27 February when they turned up to inspect the Mark III free-electron laser, which has been at Duke since 1986 and is being used to develop eye- and brain-surgery techniques. Its inventor, John Madey, moved to the University of Hawaii in 1997 and wants the device back. Hawaii officials intend to use it to develop nuclear- and chemical-weapons detectors and have already renovated space to house it.

The Department of Energy (DOE), which funded much of Madey's early work on the laser, has agreed to help get it. Last month the DOE wrote to Duke saying it was coming to inspect the 20-metre instrument as a preamble to removing it. But Duke assistant counsel Kate Hendricks told them that they would not be given access. The DOE says it will now review documents that Duke says support its claim of ownership.

Victims of smallpox vaccine 'should be compensated'

Washington The US government's smallpox-vaccination programme received a boost last week, as health secretary Tommy Thompson announced a compensation plan for any medic who is harmed by the vaccine.

President George W. Bush decided in December to vaccinate healthcare workers against smallpox, as they could be exposed to infected people if the virus were used in a terrorist attack. But hospital employees boycotted the scheme, saying the vaccine was unsafe and liability issues were unresolved. After six weeks, only 12,400 of the 500,000 intended recipients had been vaccinated.

Thompson's solution, which is yet to be approved by Congress, includes payments of two-thirds of lost wages, up to a maximum of \$50,000, and \$260,000 for family members in the event of the vaccine causing death.

US Army plans assault on biotech frontline

San Diego The US Army wants to tap the biotechnology research power of the country's universities to improve the fighting capabilities of future soldiers.

The army has invited groups of three universities to form partnerships with industry to create an Institute for Collaborative Biotechnologies, which would



The US army hopes for a boost from biotech.

receive \$37 million over five years for basic research. Sensors for biological and chemical weapons and new medical devices are just some of the hoped-for innovations.

In the past, the army has been criticized for its handling of such schemes. In 1999, for example, it awarded a \$45-million, five-year grant to the University of Southern California without taking bids from other universities (see *Nature* 400, 808; 1999). About two dozen major universities have expressed interest in the new institute.

Supernova study shows how ancients saw the light

Washington Details of an exploding star that dazzled sky-watchers as far apart as Egypt and Switzerland in 1006 have been revealed by US researchers. Chinese astronomers called it the "visiting star", and reported that after it first appeared on 1 May in that year it remained visible for months, even in daylight.

Frank Winkler, an astronomer at Middlebury College in Vermont, and colleagues have now used the Cerro Tololo Inter-American Observatory in Chile to probe the speed and velocity of the expanding shell of gas surrounding the site of the explosion.

They found that the supernova lies about 7,100 light years from Earth, and would have had a brightness between that of the Moon and Venus. "There's no doubt that this would have been a truly dazzling sight," says Winkler. "In the spring of 1006, people could probably have read at midnight by its light."

Clarification In highlighting problems in deposited sequence data of the rat genome, the editorial 'Sacrifice for the greater good?' (*Nature* 421, 875; 2003) did not make clear that such problems are common in pre-publication assemblies. We did not intend to imply that the rat sequence data are unusual in this respect.

Cosmology gets real

Cosmic insight: high-resolution plots of the Universe, such as this image showing 221,282 galaxies in a slice of sky, are helping cosmologists to confirm their theories.

2DF GALAXY REDSHIFT SURVEY TEAM

By clarifying the age and make-up of the Universe, researchers have ushered in an era of precision cosmology. Now they are preparing to probe the mysteries of dark matter and dark energy. Geoff Brumfiel reports.

The history of cosmology is full of deep thoughts and unlikely ideas. Early European star-gazers thought that Earth was at the centre of a Universe filled with concentric crystalline spheres, whereas their counterparts in America believed that the cosmos sat on the back of a giant tortoise. Much modern cosmology has been similarly speculative, with theorists using fractal patterns to describe the Universe, or modelling its evolution on the behaviour of a giant ball of very cold fluid.

Such cosmological guesswork flourished because important parameters of the Universe, such as its age, could not be measured precisely. But over the past two decades, astronomers have quietly been measuring these details with ever greater precision, verifying and discarding theories along the way. With a crucial new map just in showing the cosmic microwave background (CMB) — radiation left over from the early Universe — the field is rapidly converging on a single picture of the cosmos.

But that picture is extremely odd. A mysterious force, known as dark energy, is pushing the heavens apart. Most of the mass in the cosmos remains unseen, and researchers are unsure what form this 'dark matter' takes. The two theories that describe the Universe — quantum mechanics and general relativity — remain incompatible. So to plug these gaps in the Universe's narrative, researchers are gearing up with a new

generation of precision instruments.

Many aspects of our current picture of the Universe date from the early 1980s, which was a difficult period for astronomers. By studying the large-scale structure of the Universe — such as how galaxies are aligned — and the way in which it is expanding, researchers reached some basic, if shaky, conclusions.

Level-headed approach

Most believed that the geometry of our Universe is flat, meaning that it will continue to expand forever. For this to be true, the mass density of the Universe must have a particular value. But estimates of the visible mass in the Universe fell well short of this figure. Researchers knew that there was mass out there that they couldn't see — this dark matter revealed itself by the gravitational pull it exerted on nearby galaxies. Even so, calculations that combined the inferred dark matter with the observed luminous matter had error margins of 25–30%, making it hard to tell if the Universe really was flat.

The theorists, by contrast, were having a field day. Their ranks had been swelled by high-energy physicists, who realized that the knowledge gleaned from particle accelerators could be applied to the high-energy conditions of the early Universe. "In those days it was a theorist's paradise because there were hardly any data to go on," says Peter Coles, a theoretical astrophysicist at the University of Nottingham, UK. "You had a lot of

freedom to work on theoretical ideas without any real prospects of them being tested."

The most popular of these theories was inflation, developed in part by Alan Guth of the Massachusetts Institute of Technology. This suggests that the Universe expanded rapidly 10^{-35} seconds after the Big Bang, and it makes specific predictions about the distribution of matter and energy in the Universe immediately after this expansion. The theory says that the Universe was then a sea of high-energy particles, such as electrons and photons, which would have contained small areas of low and high density. These areas should have left their mark in the form of temperature variations in the CMB, which has had little interaction with matter since about 300,000 years after the Big Bang. By verifying these predictions, observational astronomers began to catch up with the theorists.

Temperature fluctuations in the CMB were first spotted in 1990 by NASA's Cosmic Background Explorer satellite¹, and a finer plot of these variations was provided by BOOMERANG, a small balloon-based telescope that was first flown over Antarctica² in 1998. Both maps seemed to show the temperature variation predicted by inflation. "The results had a sense of rigour that gave people confidence in where they were going," says Saul Perlmutter, an astronomer at Lawrence Berkeley National Laboratory in California.

Other measurements have since provided more support for inflation. The uneven dis-

tribution of matter in the early Universe, detected by the CMB studies, seeded the formation of the first stars. This allowed theorists to predict where and when stars and galaxies should have formed. These predictions have now been tested, thanks to two surveys — the Anglo-Australian Two-Degree Field Galaxy Redshift Survey³ and the Sloan Digital Sky Survey⁴, based at Apache Point Observatory in New Mexico — which have taken the best-ever large-scale images of the Universe. These images detailed the ways in which galaxies are aligned and provided additional information about the distribution of dark matter. The results fitted nicely with the idea of inflation and its effect on the CMB, but still left about two-thirds of the mass in the Universe unaccounted for.

In with a bang

As evidence for inflation piled up, a solution to this problem was proposed. In 1998, a group led by Perlmutter⁵ and a second team headed by Brian Schmidt⁶ of the Mount Stromlo Observatory near Canberra, Australia, measured the distance from Earth to exploding stars in far-off galaxies using two separate techniques. First they measured the colour of the light emitted by the explosions. In most cases, the farther away an object is, the redder it appears. This 'redshift' is the most common measure of distance used in modern astronomy. But astronomers also know that the explosions have very similar luminosities, so they can infer distance by measuring how bright the explosions appear.

Perlmutter and Schmidt's teams found that the brightness measurements placed the explosions, commonly called supernovae, farther away from Earth than did the redshift readings. Redshift is based on the assumption that the rate of the Universe's expansion is slowing down, so this finding led them and others to make an extraordinary suggestion: that the expansion of the Universe is accelerating, pushed outwards by some kind of phantom force for which there was no explanation.

This phenomenon of dark energy seemed

Invisible touch: the Universe's dark matter reveals its presence through gravitational lensing — the distorting effect it has on the light from galaxies.



Studying supernovae (lower left) has convinced Saul Perlmutter that cosmic expansion is accelerating.

odd. But according to the general theory of relativity, mass and energy are equivalent. And when cosmologists looked at the amount of energy needed to create the mysterious force, they found that it accounted perfectly for the mass still missing from their picture. "It's like a funny-looking jigsaw piece that just happened to fit," says Michael Turner, a cosmologist at the University of Chicago. "People were quick to accept it."

With the addition of the latest data on the CMB^{7,8}, courtesy of NASA's Wilkinson Microwave Anisotropy Probe, our picture of the Universe is now clearer than ever. Combined, the various CMB studies have confirmed that the Universe is indeed flat. The Wilkinson probe has now set ratios for the composition of the cosmos: 23% dark matter and 73% dark energy, leaving only 4% for galaxies, stars and people. The Universe's age has also been nailed to within 1% of 13.7 billion years. And the total mass density matches that predicted by inflation to within a 2% margin of error. "This is a giant leap forward in credibility for cosmology," says Max Tegmark, a cosmologist at the University of Pennsylvania in Philadelphia. "It has been transformed into a real, hard science."

Although that leap has left cosmological theory on a surer footing than ever before, the holes in our knowledge are still considerable. Researchers are confident that dark energy and dark matter are out there, but they don't know what kind of entities they are or how to find them. "Ninety-six per cent of the Universe is stuff that we've never seen," says Turner.

Hidden depths

Dark matter is the older, and perhaps easier, problem to address. The key to understanding it lies in its effects on stars and galaxies. According to general relativity, all mass distorts the space around it. When light from distant objects passes close to dark matter, it should be bent — a process called gravitational lensing. In the past, astronomers could detect only the largest clumps of dark matter, which create dramatic lensing effects. But with the aid of computer algorithms, they can now pick up much weaker distortions.

These lensing projects are only just under way, but they should produce results in the next few years, says Nicholas Kaiser, an astronomer at the University of Hawaii in Manoa. He predicts that the number of lenses identified will increase by a factor of ten. The hope is that these will allow astronomers to determine the distribution of dark matter in the present-day Universe. Those measurements won't say much about what dark matter actually is, but they should provide tighter constraints on the role it played in the evolution of the Universe.

Cosmologists also know a little about how dark matter interacts with other matter. The faster a particle moves, the more energy it transfers to any particles that it collides with. If, during the early Universe, dark matter was moving at close to the speed of light, it would have left its mark on the process by



Let's be precise

A new generation of experiments promises to provide a wealth of data against which cosmologists can test their theories. They include:

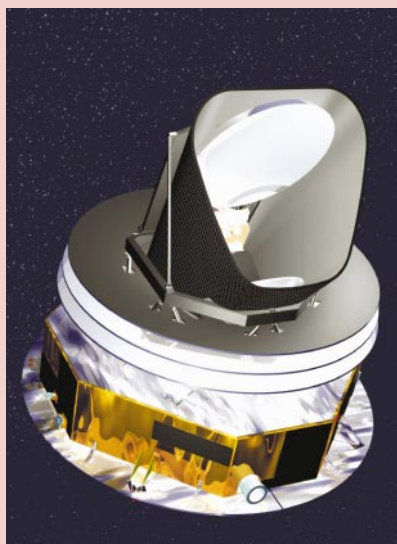
Planck A cosmic-microwave-background satellite with better spatial resolution and temperature sensitivity than the Wilkinson probe. Scheduled to launch in 2007.

James Webb Space Telescope The successor to the Hubble Space Telescope, it will allow astronomers to view younger galaxies than is currently possible. Scheduled to launch in 2011.

Supernova/Acceleration Probe Designed to take measurements of distant supernovae to see if the strength of dark energy has changed over time.

Large Synoptic Survey Telescope A giant ground-based telescope that would improve maps of large-scale galactic structure and probe dark matter using gravitational-lensing techniques.

Laser Interferometer Space Antenna Three satellites that would look for gravitational waves generated by inflationary expansion, an important but unchecked prediction of the inflation theory.



Planck: set to improve our picture of the cosmos.

which matter clumped together to form stars and galaxies. But astronomers can watch star and galaxy formation occurring in very distant parts of the Universe, and so far they have not seen any evidence of the influence of fast-moving dark matter.

This has led many cosmologists to speculate that it is made up of heavy and relatively slow-moving particles that seldom interact with visible matter. This prescription interests high-energy physicists, who may be able to help. Although nothing that fits the bill has yet been created in a particle accelerator, researchers at CERN, the European Laboratory for Particle Physics near Geneva, should have a new and more powerful device — the Large Hadron Collider — up and running in 2007. Theoretical calculations indicate that the collider may create candidate dark-matter particles, which would help to constrain cosmologists' models.

Dark energy is a more vexing problem, but the solution could lie in the nature of empty space. According to quantum theory, particles and their antiparticle equivalents are continually being created and annihilated, even in a vacuum. Some researchers have speculated that this vacuum energy could be what is accelerating the Universe's expansion. But theoretical predictions for vacuum energy are up to 120 orders of magnitude greater than the strength of dark energy seen today.

Another idea, says Paul Steinhardt, a theoretical physicist at Princeton University in New Jersey and one of the originators of inflation theory, is that dark energy may be a cousin of inflation. Like dark energy, the inflationary force resisted gravity, pushing everything outwards. But inflation was many times the strength of the current outward-acting force, and seemed to switch off moments

Ninety-six per cent of the Universe is stuff that we've never seen.

after the Big Bang. Theorists are now experimenting with weaker versions of the equations for inflation to see if they can describe dark energy. So far, few solid conclusions have emerged. "Those models raise more questions than they answer," says Turner.

Highly strung

Other theoretical ideas may come from attempts to combine the worlds of quantum mechanics and general relativity. Known broadly as string theory, these models suggest that all elementary particles are made of tiny strings and loops vibrating in multiple dimensions. The theory might be able to explain how tiny quantum fluctuations, such as vacuum energy, could interact with gravity on larger scales to create the effect of dark energy. "String theory might give you answers," says Lisa Randall, a theorist at Harvard University who is working on a version of inflation that involves extra dimensions. "But I don't think anything has come close to doing it yet."

As in the past, experimentalists might eventually provide the additional data needed to constrain these models. In 2001, Adam Riess, an astronomer at the Space Telescope Science Institute in Baltimore, Maryland, led a Hubble Space Telescope team that helped to confirm dark energy's existence by looking at older supernovae⁹ than those studied by Perlmutter and Schmidt. Riess is now

looking for even older supernovae to see if dark energy changes with time, although his results will only be able to detect large deviations in the force's behaviour.

Apart from supernovae, there are no other bright and distant objects that lend themselves to distance measurements, so astronomers are wracking their brains for new methods of probing dark energy. One of the most promising is the Deep Extragalactic Evolutionary Probe 2 at the Keck Observatory on Mauna Kea in Hawaii. The study is trying to establish what types of object — most probably certain kinds of galaxy — are evenly distributed throughout the Universe. By monitoring collections of these objects at different distances from Earth, and hence at different stages in the evolution of the Universe, the survey will reveal how dark energy pushes them apart, which should help to verify the supernovae results.

A new generation of precision instruments, planned for the next decade and beyond, should provide a slew of extra data (see 'Let's be precise', left). The Supernova/Acceleration Probe (SNAP) satellite would, for example, continually monitor patches of sky, searching for new supernovae. The probe's backers, based across several US institutions, say it could gather data on 2,000 supernovae a year — 20 times the number obtained by a decade of ground-based searches.

Such data would allow astronomers to tell whether dark energy is constant, as the vacuum-energy explanation would suggest, or changing, as its inflationary cousin did in the early Universe. But researchers working on SNAP are still seeking funding for the project, and the probe is unlikely to be launched before the end of the decade. "It could take us a long time to figure out the nature of the dark energy," says Steinhardt.

For many astronomers, the CMB map generated by the Wilkinson probe is a milestone in cosmology. The data capped a long effort to measure the basic properties of the Universe. And despite the unexplained phenomena of dark matter and energy, astronomers are more confident than ever of their ability to understand the cosmos. "We've flushed out the basic features of the Universe," says Turner. "What we need now is a good story."

Geoff Brumfiel is *Nature's* physical sciences correspondent in Washington.

1. Smoot, G. F. et al. *Astrophys. J.* **396**, L1–L5 (1992).
2. de Bernardis, P. et al. *Nature* **404**, 955–959 (2000).
3. Percival, W. J. et al. *Mon. Not. R. Astron. Soc.* **327**, 1297–1306 (2001).
4. Zehavi, I. et al. *Astrophys. J.* **571**, 172–190 (2002).
5. Perlmutter, S. et al. *Nature* **391**, 51–54 (1998).
6. Garnavich, P. M. et al. *Astrophys. J.* **493**, L53–L57 (1998).
7. Reichardt, T. *Nature* **421**, 777 (2003).
8. Carroll, S. *Nature* **422**, 26–27 (2003).
9. Riess, A. G. et al. *Astrophys. J.* **560**, 49–71 (2001).

Wilkinson Microwave Anisotropy Probe
map.gsfc.nasa.gov

Against the grain

China has long been a keen supporter of transgenic agriculture, and is still pouring money into developing the technology. So why are applications to market new genetically modified crops in limbo? Colin Macilwain investigates.

Last June, the *China Daily* — the English-language mouthpiece of the country's ruling Communist Party — published an article by Greenpeace, describing the alleged ecological risks posed by transgenic crops. It wasn't quite as though Amnesty International had been asked to write a piece on China's treatment of its political dissidents, but for those who are familiar with Beijing's official line on transgenic agriculture, it still marked a dramatic turnaround.

Since the mid-1980s, the Chinese government has ploughed hundreds of millions of dollars into developing the technology, and in the late 1990s it swiftly authorized the commercialization of a handful of transgenic crops. At that time, Chinese agriculture seemed to be on the highway to a genetically modified future. Indeed, at the lab bench, that is still the case: universities and government labs are bursting with ideas, talent and investment. "Plant science is a major activity, because it is so important to China," says Jiayang Li, director of the Chinese Academy of Sciences' Institute of Genetics and Developmental Biology in Beijing.

But in the past couple of years, agricultural biotechnology in the world's most populous nation has taken a detour into a cul-de-sac. No new transgenic food crop has been approved for commercial use since 2000 — although candidate crops continue to move into field trials. And given signals such as the Greenpeace article, the outlook for approvals seems to be worsening. "My personal view is that the current discussion is biased against agricultural biotechnology," says Zhixue Wang, the official in charge of the Ministry of Science and Technology's Rural Technology Development Centre in Beijing.

What's going on? It is hard, in a country that only allows free debate within carefully prescribed bounds, to put your finger on the answer. Chinese agribiotech researchers cite "public concerns" about biosafety, as well as doubts that export markets in Europe and elsewhere will accept genetically modified produce. But circumstances suggest an alternative explanation: that the Chinese government is exploiting the biosafety issue to frustrate the commercial ambitions of Western agribiotech firms, because it realizes that its own research



Sowing the seeds: China's farmers (right) currently prefer high-yielding imported transgenic cotton to homegrown varieties. Anxious to avoid a similar situation prevailing for rice (above), the government is now investing heavily in research on the crop.

programme needs more time to catch up.

The Greenpeace article, like much of the *China Daily*'s contents, carries a coded message. And in this case, that message seems to be: "Monsanto, keep out." The world's leading agribiotech firm, based in St Louis, Missouri, has already carved out a sizeable share of China's market for cotton seed, selling varieties that are engineered to produce an insecticidal toxin. And China's farmers — or at least those who can afford it — are prepared to pay a sizeable premium for the high-yielding Monsanto products, in preference to cheaper homegrown varieties.

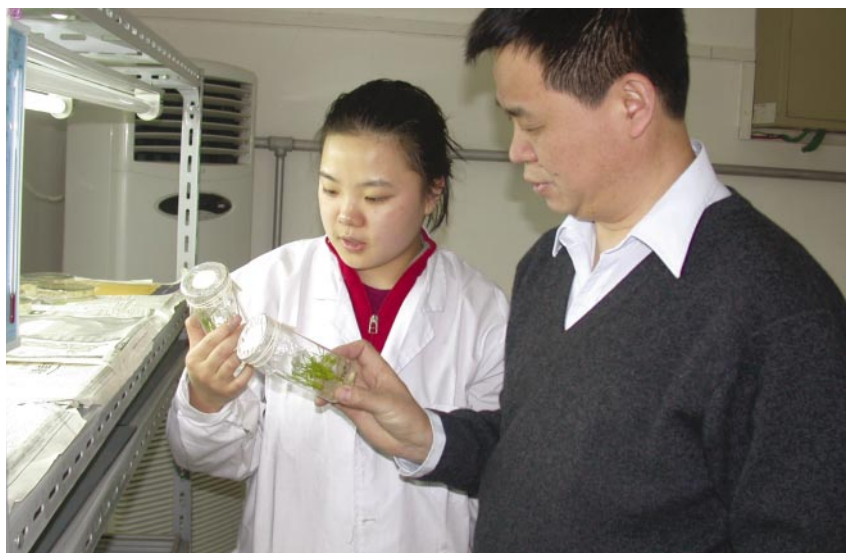
Cornering the market

In the circumstances, say some Western observers, it makes sense for Beijing to close the door on new commercial approvals until its domestic products can compete effectively. In the meantime, at least viewed from an American perspective, China is using European public concerns about the safety of

transgenic crops to keep imported varieties at bay. "China is trying to make major investments in biotechnology research," says Julia Moore of the Smithsonian Institution's Woodrow Wilson International Center for Scholars in Washington DC, who follows the global trade in transgenic crops. "But it is also taking advantage of biotechnology concerns in Europe and elsewhere to limit its imports of the technology."

Given the scientific effort that is currently under way in Beijing, one would hardly think that this is a country that is putting the brakes on transgenic agriculture. Research laboratories run by the Chinese Academy of Sciences and the Chinese Academy of Agricultural Sciences (CAAS), plus those at Peking University and the China Agricultural University, are well equipped and generously staffed. At Li's institute, for example, some 25 principal investigators and 300 other staff and students are engaged in plant science, mostly related to wheat and rice.

STILL PICTURES; INSET, A. WRIGHT/PANOS PICTURES



Jiayang Li (right) is unimpressed with agribiotech firms' reluctance to engage with Chinese researchers.

China has a long-standing relationship with transgenic technology: as far back as 1993, its farmers were growing genetically modified tobacco that was resistant to insect attack. But when Philip Morris, the US tobacco company, heard farmers boasting about the crop, they barred its purchase for use in their cigarettes. That episode, several researchers say, led to establishment of a formal approval system for genetically modified crops.

Under the system, transgenic crops that have been approved for sale include a tomato modified for increased shelf-life, and sweet peppers resistant to cauliflower mosaic virus. But in economic terms, by far the most important are several varieties of 'Bt' cotton, which produce an insecticidal toxin derived from the bacterium *Bacillus thuringiensis*.

In the early 1990s, almost one-third of China's vast cotton crop was being lost to bollworm, the voracious caterpillar of the moth *Helicoverpa armigera*. In 1997, the government approved commercial planting of a variety of Bt cotton developed by Monsanto and designed to combat the pest. Since then, more Bt varieties, some from Monsanto, others produced by the CAAS and other domestic labs, have also won approval.

According to one survey published last year, farmers growing Bt varieties in 1999 reported yields that were 6% higher and production costs some 28% lower than farmers cultivating conventional cotton. Most strikingly, the transgenic farmers had slashed their use of synthetic pesticides by more than 80% (J. Huang, S. Rozelle, C. Pray and Q. Wang *Science* **295**, 674–676; 2002).

China's own Bt cotton varieties are subtly different from Monsanto's versions. Most Bt crops contain only one of two related *B. thuringiensis* genes, called *cry1Ab* and *cry1Ac*, but the Chinese varieties include both. The genes were also introduced using

a different method. Like most transgenic crops, Monsanto's Bt cotton was made by bombarding plant tissue cultures with tiny particles of tungsten or gold coated with DNA containing the transgenes — a process patented by the seed company Pioneer Hi-Bred of Des Moines, Iowa. But the Chinese varieties were created using a different technique, which has not been published in the international literature but was patented in 1998. The method involves injecting DNA into the seed embryo in the plant ovary — through the tube created by the pollen that fertilized the ovum — within 24 hours of the cotton flowering.

Homegrown talent

This go-it-alone approach is a source of considerable pride for Chinese researchers. "China was the second country in the world to develop a genetically modified crop and get intellectual property rights to it," boasts Dafang Huang, director of the CAAS's Biotechnology Research Institute in Beijing. Today, 0.6 million hectares of the total Chinese crop of 4 million hectares consists of the homespun transgenic cotton.

Monsanto's products, however, account for a similar cultivated area — even though its seed is up to ten times costlier. The reason that farmers are prepared to pay a premium for the imported varieties, Chinese researchers admit, is their superior quality.

One solution would be for Chinese researchers to collaborate with Monsanto to address the country's agricultural needs. "In 1999 I held talks with the president of Monsanto in the Far East," recalls Wang, "and we hoped to use their expertise to develop crop varieties here in China." But neither Monsanto nor the Beijing government is a shrinking violet in the field of intellectual property negotiations, and the talks reached deadlock — leaving the company with near-pariah

status among Chinese scientists. "Monsanto seems to want its products in China, but not its research," sniffs Li.

Li and other institute heads speak more warmly of Swiss-based Syngenta and Pioneer Hi-Bred, both of which have set up several collaborations with individual research groups. Even still, no Western firm has yet established a broad-based collaborative research agreement with a Chinese institution.

Most observers agree that the current hiatus in commercial approvals for transgenic crops is likely to deter Western agribiotech companies from broadening their links with Chinese labs. Such approvals are the responsibility of a subcommittee of the National Biosafety Committee, which meets twice a year. Its meetings are closed, and no account of its workings has been published since the brakes were put on new approvals. "At the moment it is difficult to get approved," says Guoying Wang, head of the molecular-biology department at the China Agricultural University and a member of the subcommittee, rather vaguely. "Most scientists support biotechnology but some scientists are opposed, including some senior scientists."

The subcommittee's decisions, at least in theory, are underpinned by a biosafety research programme that was launched in 2001. According to Yufa Peng, a plant pathologist at the Institute of Plant Protection in Beijing and the CAAS's chief scientist on the biosafety programme, the effort involves 125 researchers at 23 laboratories. It includes research into the ecological impact of transgene flow to crops' wild relatives, and the effects of transgenic agriculture on biodiversity.

Beijing's official policy, meanwhile, is deliberately equivocal. As Liu Xu, vice-president of the CAAS, puts it: "China will abide by the precautionary principle and substantial equivalence." The former is usually invoked to reject transgenic technology on grounds of caution; the latter implies acceptance on the grounds that transgenic crops are not significantly different to those produced by conventional breeding.

Conveniently enough, this ambiguity leaves China free to embrace the commercialization of transgenic crops once more, when it feels that its own technology can compete with Western imports. If this is the real reason for Beijing's current position, then the progress of China's research on rice may be the crucial factor. Already, Chinese researchers have published a draft genome sequence for the *indica* subspecies of rice (J. Yu *et al. Science* **296**, 79–92; 2002). The government is now investing heavily in rice functional genomics and in efforts to genetically engineer the crop. "Chinese institutions still have a lot to learn," says Huang. "But hopefully the situation with rice will be better than it was with cotton."

Colin Macilwain is *Nature's* news editor.

Work on 'non-lethal' weapons should be limited too

Security fears have led to violation of international law before. It must not happen again.

Sir — Given the prevailing threat of terrorist use of biological weapons, I welcome the statement by journal editors and authors (*Nature* **421**, 771; 2003) affirming their responsibility to censor work that could compromise biodefence and biosecurity. This measure may help avert misapplication of research findings.

However, the threat posed by 'non-lethal' weapons research, the results of which are not usually submitted for publication, also warrants attention. These include unpublished research into weapons such as 'calmative gases' and biologically engineered anti-material bacteria.

Given the casualties they can claim (see *Nature* **420**, 7; 2002), the term 'non-lethal' weapons is a misnomer. Research on such weapons might violate international chemical and biological anti-proliferation treaties. Just as journals and authors are taking a stand against publishing potentially dangerous research findings, so too should the world's scientists actively condemn proactive

research into these kinds of weapons.

The South African experience has shown how an acquiescent attitude towards weapons research that violates international law, on the grounds of patriotism and security fears, can compromise the legal and ethical responsibilities of people and research institutions. From 1983 to 1993, under the auspices of Project Coast — the apartheid government's covert chemical and biological weapons programme — South African scientists, physicians and academic experts put their expertise into developing secret and novel weapons (see *Nature* **393**, 724; 1998). In so doing, they unleashed agents and weaponry that violated South Africa's international-law obligations.

History now questions the ethics of the scientists who contributed to this programme, and of the institutions to which they belonged. History will judge today's scientists and institutions similarly should they turn from their ethical and legal obligations in the face of security pressures.

I do not suggest that scientists should

totally refrain from involvement in weapons research. But, for the sake of world security, scientists in all disciplines that could conceivably be involved in novel 'non-lethal' weapons development should learn from the South African experience.

They must commit themselves to knowing, respecting and adhering to the legal instruments that govern chemical and biological warfare. They must commit themselves to drafting a scientific moral code based on benevolent ethical principles. They must lobby their governments to adhere to anti-proliferation laws and declarations, and to ratify verification protocols associated with these declarations. They must report infringements of these laws to the relevant authorities. Most importantly, they must ensure that a Pandora's box of deadly 'non-lethal' weapons never gets built, let alone opened.

Jerome Amir Singh

Howard College School of Law, University of Natal, Durban 4041, South Africa, and University of Toronto Joint Centre for Bioethics, 88 College Street, Toronto M5G 1L4, Canada

Universities could gain from backing biotech

Sir — Since the inception of the biotechnology industry more than two decades ago, the major pharmaceutical companies have supported discovery-phase research in the sector through collaborations and out-sourcing. Even when the economics seemed unfavourable to the biotechnology company, these arrangements have been beneficial in critical ways.

But the earliest phases of R&D — target discovery, validation and screening — are receiving less and less support within the pharmaceutical industry. The major companies are under enormous pressure to continue high growth, while dealing with unspectacular results of genomics collaborations, consolidation and the establishment of internal infrastructure. All this has greatly reduced their appetite for discovery-phase partnerships with biotech start-ups.

A preference for late-stage product opportunities will create long-term problems for the drug industry and for health care in general. Clinical compounds do not materialize out of thin air. If the pipeline input is squelched, the downstream output will eventually tail off.

The change in the commercial funding environment creates some interesting

opportunities for academic institutions. Without departing from their primary mission of assisting the public good through education and other services, they could seize the chance to fill this gap and attract government funds for discovery-stage work. As projects mature, universities could create mechanisms to transfer programmes into a commercial setting where they can exploit greater amounts of private capital and command higher returns on investment than has typically accrued to them.

Universities can already tap into large sums of public money to support work of medical value. If they could undertake a focused effort on early-stage technology and drug-discovery R&D, I believe that a relatively simple and productive transfer into the private sector could be arranged. The universities could find the business expertise necessary to launch companies successfully through their internal resources and through cultivated venture-capital groups. Academic institutions could thus exert considerable control over commercialization of their discoveries.

A properly conceived and managed university unit of this type would require an initial investment of funds before becoming self-sufficient. Once the infrastructure is built, a unit could rapidly become a source of income for the university through government grants,

research contracts and, ultimately, its own commercial ventures.

Returns on the initial investment would probably derive in part from royalties on exclusive (for example, drug) or non-exclusive (for example, technology) licences. In addition, universities could gain income through equity positions in companies that they help to found, and institute regulations to ensure openness and mitigate perceptions of conflict and exploitation.

Universities stand to benefit in several other ways. Much of the infrastructure for target validation and drug screening operations could be used to accelerate, and possibly transform, their own applied or basic research activities. These services could be provided by the unit as a core facility or as collaborative projects. For instance, a target validation or chemical screening system could be used by any faculty member wishing to explore the roles of his or her favourite genes in a specific disease area.

Alexander Kamb

Deltagen Proteomics, 615 Arapsee Drive, Salt Lake City, Utah 84108, USA

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

Constellations in a cellular universe

We must agree now on strategies to search the proteome.

Ruedi Aebersold

Proteomics — the systematic analysis of proteins — has raised grand expectations for biology and medicine, exceeding those of the other genomics technologies. Yet these expectations will be difficult to realize unless proteomic technologies are disseminated beyond specialized laboratories, and made robust. A substantial shift in focus is needed, and here I suggest how this might be accomplished using a novel browsing technology. Once adopted, successfully implemented and supported, this technology has the potential to achieve these goals.

Success with genes

One of the most striking results obtained from completed genome sequencing projects is the knowledge of the precise number of genes in the genome of a species. Although the number of genes analysed to date is relatively small — ranging from a few hundred for bacteria to tens of thousands for mammalian species — the number of possible products encoded by these genes is much higher. In particular, the number of encoded proteins is enormous, as the same gene can generate multiple protein products that differ as a result of combinatorial splicing, processing and modification.

As the universe of a species' biological processes and the molecules that constitute these processes is finite and knowable, there are several important consequences for experimental biology. First, even the most complex biological phenomena such as development, differentiation, metabolism and memory will be explained using known genes, their products and their interplay with the external conditions that the organism encounters. Second, projects to discover genes and their products in a species have defined end points. So far, although this end point has been reached only for gene discovery (sequencing), at some point in the future all of a species' gene products — including messenger RNA and proteins — will also be comprehensively described. Third, once all the possible molecules and activities within a species have been discovered and described, biological experimentation will be transformed from a discovery mode of identifying and describing molecules, to a 'browsing' mode, in which the universe of possible events is searched to find constellations that correlate with a particular state or function. Genomics-style biology can therefore be separated into two distinct phases: a discov-



Mapping of inner space: proteomics could provide a unique perspective of how cells function

ery phase to characterize the universe; and a browsing phase, in which system-wide biological assays navigate the universe.

Proteins are involved in all biological processes and can therefore be considered the functionally most important biological molecules. They are also particularly rich in biological information. In addition to the amino-acid sequence defining a protein, protein properties such as the amount of a protein expressed, its specific activity, state of modification and association with other proteins or molecules of different types are crucial for the description of biological systems. The systematic identification and characterization of proteins, called proteomics, carries with it huge expectations, such as diagnostic and prognostic markers in blood serum and other body fluids; targets for pharmaceutical drugs; and improving the knowledge of fundamental biological processes. Hence the development of technologies to search the 'proteome' routinely and systematically would be a significant achievement.

Unfortunately, the same properties that make proteins information-rich also significantly complicate their experimental analysis. There is no experimental platform, even under development, to systematically measure the diverse properties of proteins at high throughput. Currently, the most

mature and versatile proteomic methods are based on mass spectrometry. For the study of complex protein mixtures, most methods revolve around the following outline. Sample proteins are proteolytically cleaved into smaller peptides. The resultant peptides are separated and analysed in a mass spectrometer, the data processed through a series of computer algorithms that determine the sequence identity of the proteins, and to some extent their state of modification. Mass-spectrometry-based analyses are extremely accurate quantitatively, provided that suitable reference standards are available. These standards are usually generated by labelling a reference sample at a known concentration with stable isotopes, generating pairs of molecules that are chemically identical but of different mass, so that they can be distinguished by a mass spectrometer. Thus the signal intensity ratio observed between the known standard and unknown sample versions of the same molecules provide an accurate measure of the abundance of each species in the original sample.

As this rather complex process requires expensive instrumentation, information-technology infrastructure and highly specialized personnel, it is unlikely that it will spread widely from specialized proteomics laboratories to biological and clinical research communities. There are two other

R. NIGHTLEY/SPL

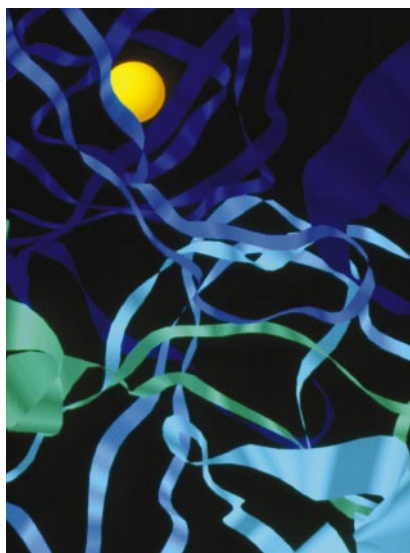
factors that make it highly doubtful that even large, specialized proteomics laboratories will achieve the automation and sample throughput that is required for undertaking huge (for example, population-based clinical) studies. First is the enormous complexity of the proteome—the human genome is assumed to contain around 30,000 genes, but it has been estimated that human serum alone contains in excess of 500,000 different protein species. Second, current mass-spectrometry-based proteomics operates in discovery mode, which requires the re-discovery of each protein species in each experiment. It is therefore essential to develop a robust proteome-browsing technology to meet the high expectations.

Current applications

In some areas of molecular biology research, the transition from a discovery to a browsing mode of operation is already in progress, notably in gene-expression array analysis and in the analysis of single nucleotide polymorphisms (SNPs). In the former, the discovery phase establishes the transcripts potentially generated by a genome (transcriptome) using experimental and computational methods. In the browsing phase, probes specific for each transcript are arranged in ordered microarrays; the transcripts of the genes expressed by a cell in a specific state are concurrently identified and quantified in a single experiment.

Comparisons of the transcript patterns obtained from cells representing different states are expected to provide detailed diagnostic patterns classifying cellular or pathological states and provide new insights into the function and control of biological processes. This method has already yielded biologically and clinically significant results. For example, transcript profiling has been used for predicting survival in breast cancer; for predicting clinical outcomes in breast cancer and survival after chemotherapy for diffuse, large B-cell lymphoma; and concurrently to monitor hundreds of cellular functions. In the discovery phase of SNP analysis, an international consortium of scientists is generating a comprehensive catalogue of this universe in the human population. In the browsing mode, called 'scoring', associations of specific patterns of SNPs with specific phenotypes are being established. Such patterns are thought to be useful for predicting the propensity for the onset of a variety of disease conditions.

For both gene-expression-array and SNP analyses, successful studies in the browsing mode were carried out before the discovery phase was complete. High-throughput analysis is inconceivable if the objects studied have to be discovered *de novo* each time. Segregation of the process into a discovery phase followed by a browsing phase was vital for the implementation of these two power-



Proteomics could allow many complex interactions to be deciphered.

ful genomic technologies and will be even more so if proteomics is to succeed.

My proposal for a browsing technology is conceptually simple. For each protein, protein isoform or specifically modified form of a protein, a peptide sequence that uniquely identifies that polypeptide is selected, chemically synthesized and labelled with tags of a heavy stable isotope. These peptides are the definitive markers for the proteins to be studied. Precisely measured amounts of these reference peptides are then added to a sample in which the proteins or peptides have been labelled with tags of a light stable isotope. The combined peptide sample can be separated reproducibly, and fractions deposited on the sample plate of a mass spectrometer, effectively generating an ordered peptide array. Each array element can then be interrogated by a mass spectrometer and will generate two types of signal, one representing the signals of the peptides for which no reference peptide has been added, appearing as single peaks, and the other representing the signals for those peptides for which a reference peptide was added, appearing as paired signals with a mass difference that precisely corresponds to the mass differential encoded in the stable isotope tag. In this method, a protein is identified by correlating the position and the accurately measured mass of each isotope-peptide pair in the array. Proteins are quantified by determining the ratio of the size of the signal of a peptide derived from the protein mixture with the signal of the corresponding reference peptide.

There are several advantages of this proteome-browsing method. First, one peptide is sufficient for the unambiguous identification and quantification of each protein. Therefore, the number of peptides that need to be analysed to identify and

quantify the product of every gene approaches the number of genes in a genome. Second, data analysis becomes trivial as each protein is identified and quantified by correlating the acquired data with a look-up table, rather than by *de novo* identification. Third, the method is easily standardized between laboratories. Fourth, the absolute quantity of each protein is determined, making data sets easily comparable. Fifth, any subset of proteins, for example proteins contained in organelles, subcellular fractions or differentiated cells, can be selectively interrogated. Sixth, splice isoforms, differentially modified or processed proteins, can similarly be absolutely quantified, provided that appropriate reference peptides can be synthesized. Finally, the method is relatively cheap, as only minuscule (nanogram to sub-nanogram) amounts of the peptide standards are used per assay. However, similar to other genomics technologies, the proposed proteomics technology requires a sizable initial cost and labour investment that will pay dividends through the wide dissemination of a rapid, robust and simple quantitative technology.

An initial investment is required for the synthesis and calibration of thousands of isotopically labelled peptides. A project of this scope not only exceeds the ambitions of a typical laboratory, but mandates a collaborative approach, for if different research groups were to generate their own sets of reference peptides, correlation of data between studies would be difficult. I therefore call for a community-based effort or a private-public partnership for the design, synthesis and distribution of suitable sets of reference peptides. Emulating the process pioneered by the yeast community for the design of polymerase chain reaction probes for each gene, a set of unique peptide standards for each yeast protein could be generated at an estimated cost of \$1–2 million and made generally accessible for the systematic study of yeast biology.

In the arena of clinical research, the quantitative profiling of serum proteins lends itself particularly well as an early application of the technology. Serum protein profiling holds enormous promise as a tool for the early detection and stratified diagnosis of disease and for the assessment of the success of therapeutic regimens. In such studies, there are potentially thousands of samples that need to be profiled using reproducible and transparent methods of exquisite quantitative accuracy, attributes that match those of the proposed method. By means of a robust proteome-browsing technology, proteomics will be able to fulfil its potential. ■

Ruedi Aebersold is at the Department of Molecular Biology, University of Washington, Seattle, Washington, 98195-7730, USA.

Move to the rhythm

Many systems, from pendulums to fireflies, work in synchrony.

Sync: The Emerging Science of Spontaneous Order

by Steven Strogatz

Hyperion/Allen Lane: 2003. 338/352 pp.

\$24.95/£14.99

J. J. Collins

One of the central challenges facing many scientific disciplines, including physiology, molecular biology and physics, is to understand the dynamics of complex networks of interconnected components. Attempts to identify generic behaviour or universal organizing principles for such systems often leads to results that are suggestive and intriguing but lack a rigorous, mechanistic foundation. Exceptions to this are studies that deal with synchrony, a phenomenon in which the dynamics of multiple components become locked in time. Synchrony has been the object of rigorous scientific enquiry since 1665, when Christiaan Huygens noticed from his sick bed that the two pendulum clocks hanging next to one another in his room were synchronized.

Steven Strogatz, a pioneer and researcher into synchronization, offers an expert tour of this area in his new book, *Sync*. His aim is to provide the reader with a sense of the diverse forms and implications of synchrony in animate and inanimate systems, and an idea of the mathematics used to model such systems. He succeeds admirably in a work that is part general science book and part memoir.

General science books typically fall into one of two categories: either they offer a solid but dry description of the science, or they present an entertaining but superficial treatment of the subject, leaving the reader craving more substance. Strogatz combines the best qualities of these two: he is a first-rate storyteller and an even better teacher. Throughout the book he makes effective use of colourful analogies to build the reader's intuition about the system or model under study. These include comparing a cardiac pacemaker cell to a leaky toilet bowl, and a boson to the character Pig Pen from the *Peanuts* comic strip. Teachers of undergraduate science and engineering courses who struggle to find ways to give students a real 'feel' for dynamics and underlying mechanisms will find much inspiration in this book.

Strogatz covers a considerable amount of material, including synchronized fireflies, chemical oscillations, superconductors, lasers, synchronized chaos and small-world networks.

My favourite stop on his tour is the chapter on "Sleep and the daily struggle



Connected: Steven Strogatz sees parallels between bosons and the *Peanuts* character Pig Pen (inset).

for sync". Strogatz explains, among other things, how Chuck Czeisler studied human subjects over extended periods of time isolation, and discovered that the sleep-wave cycle can become desynchronized from the body temperature cycle, and yet remain linked to it in a consistent way. In particular, Czeisler found that subjects who fall asleep when their body temperature is low tend to sleep for short periods, whereas those who fall asleep when their temperature is high will sleep for long periods. Moreover, the relationship between sleep duration and sleep onset time relative to the temperature cycle is discontinuous. Several hours after the body temperature reaches its low point, sleep duration jumps abruptly from very short periods to very long ones, then slowly decreases back to short durations. This gradual descent leads to the surprising effect that if you stay awake longer than usual and go to sleep later in your body's temperature cycle, you tend to sleep less, even though you may be exhausted from a late-night party.

The memoir portions of the book offer fascinating glimpses into the daily life of a scientific researcher that are some of the best I have read. Strogatz illustrates through several anecdotes the fits, starts and strange

turns that are part of cutting-edge research projects. He also indirectly emphasizes the importance of 'play' in research, the need to be bold and open to new possibilities, and the critical value of having like-minded collaborators with different but complementary backgrounds.

I particularly liked the story of how Strogatz, as a graduate student at Cambridge University, dispirited by the cultural divide between England and America, stumbled into a local bookstore and came across Art Winfree's classic book *The Geometry of Biological Time* (Springer, 1980). Strogatz tells how he revisited the shop several times to continue sampling Winfree's book, finally dipping into his Marshall Scholar stipend to buy a copy. A short time later he got up the courage to send off a letter to Winfree, which led to a gracious reply and eventually an invitation to join Winfree's lab as a summer student. This serendipitous course of events led to the pairing of two of the most creative biomathematicians of the past few decades, and yielded landmark results on scroll waves. *Sync* is filled with many such stories and colourful portraits of leading synchrony researchers, including Charlie Peskin, Nancy Kopell, Kurt Wiesenfeld and Lou Pecora.

Sync is a great read. I will recommend it

CERN:INSET: UNITED FEATURE SYNDICATE

to my students who are considering graduate studies and possible careers in research, as well as to friends and colleagues who are looking for a starting point to explore and understand the dynamics of natural and technological networks. ■

J. J. Collins is in the Center for BioDynamics and the Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA.

All in the mind?

Placebo: The Belief Effect

by Dylan Evans

HarperCollins: 2003. 240 pp. £16.99

Walter A. Brown

In recent years the placebo response has captivated both the scientific community and the general public. *The Lancet* ran a series of articles on the placebo effect in 1994, and the US National Institutes of Health (NIH) sponsored a conference on it in 2000. Over the past five years, television programmes, newspapers and news magazines from around the world have carried stories about the placebo effect. Why the recent flurry of interest in a phenomenon that has been recognized for most of the past century and subjected to a low but steady rate of enquiry for 50 years?

Part of the fascination comes from the recent attention given to the mind-body relationship. Technology-driven advances in our knowledge of the brain and an interest in alternative and complementary medicine, with its emphasis on self-healing and spirituality, both focus attention on the links between mind, brain and illness, the arena in which placebo operates. The high visibility of clinical trials in recent years has also drawn attention to the placebo response. These trials keep telling us that for some conditions, such as depression and pain, placebo treatment brings almost as much benefit as 'real' drugs and surgical procedures. They have also spotlighted the ethics of placebo controls.

The placebo effect has long been considered a nuisance that obscured the identification of new treatments. For most of the past century, in the rare instances when doctors deliberately gave patients placebos, they did so to placate patients who were thought to be complaining too much, or to 'prove' that a symptom was psychologically based rather than 'real'.

But these views are changing. The placebo response is now widely regarded as a phenomenon worthy of study in its own right. After its placebo conference, for example, the NIH requested grant proposals to study both the basic mechanisms and the clinical application of the placebo response. Recent studies have shown that functional changes in the brain are associated with the placebo

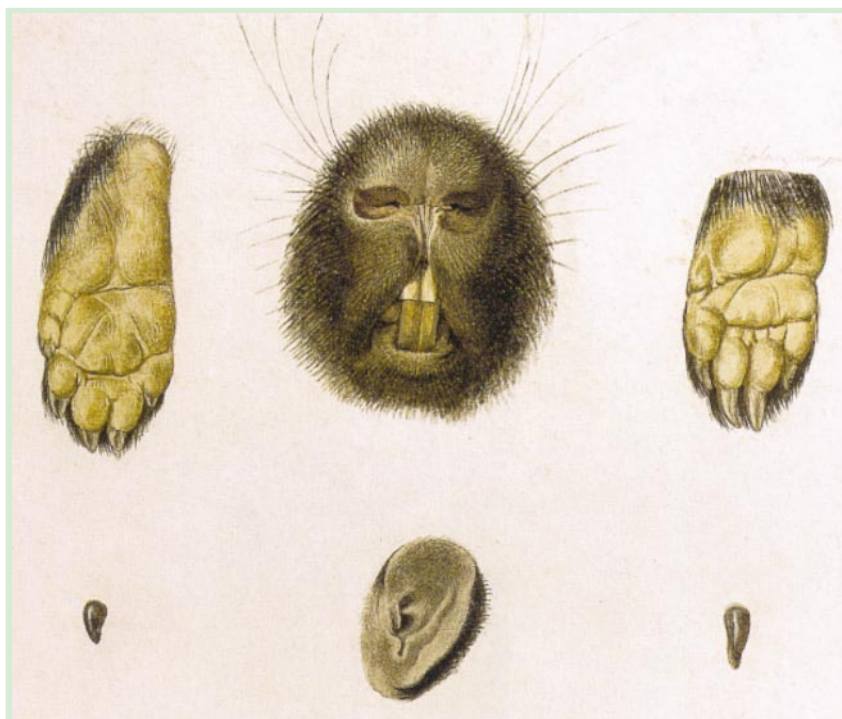
response in both depression and Parkinson's disease. And some clinicians are beginning to consider the placebo response as an integral part of healing, and are seeking ways to harness and enhance it.

Given the subtitle of Dylan Evans' book, one might expect it to offer a comprehensive assessment of the role of belief in the placebo effect. But that's not what we get. Instead there is a smorgasbord of facts and speculation presented in support of the author's main — and in my view thoroughly implausible — hypothesis that "all placebo-responsive conditions" involve activation of the inflammatory response and that placebos work by suppressing inflammation. If a highly placebo-responsive condition such as depression doesn't seem to involve the inflammatory response, Evans is undeterred: he comes up with a bit of information (in this instance one study showing that depressed patients have increased interleukin levels) to convince us otherwise. He manages to fit Parkinson's disease into his model, along with a raft of other ills that few others would ever consider as diseases of inflammation. Evans does state at the outset that this is an unsubstantiated hypothesis,

but later in the book, as he gets rolling, he seems to assume that he has demonstrated its validity.

His case is not helped by inaccuracies ranging from trivial errors of fact (he tells us that no study of depressed patients has included both a placebo and a no-treatment group, but there is at least one such study, and it is cited in a paper he reviews in depth), to misinformation (he says that placebos don't help asthma) and major misconceptions (that classical conditioning entails belief in the connection between the conditioned and unconditioned stimuli).

Evans does provide a reasonable critique of the widely discussed meta-analysis by Asbjørn Hróbjartsson and Peter Gøtzsche (*New England Journal of Medicine* **344**, 1594–1602; 2001) that compared placebo to no treatment. The data were largely misinterpreted, including by the authors, as showing that a placebo offers no more benefit than the passage of time. And Evans includes an entertaining chapter speculating on the evolution of the placebo response and how the capacity for it might be adaptive. It's utter nonsense, but I'm a natural-selection junkie and I enjoyed the attempt.



Through the eye of the lynx

Italy is brimming with art treasures, but now and again new troves emerge from obscurity. In *The Eye of the Lynx* (University of Chicago Press, \$50), art historian David Freedberg brings together a wealth of mostly unknown natural-history drawings produced by the world's first modern scientific academy, the Accademia Nazionale dei Lincei (Academy of the Lynxes),

which was founded 400 years ago. Revolutionary at the time, the drawings were aided by the use of a new instrument — the microscope — which was created by turning around Galileo's new telescope. Pictured here is a selection of Vincenzo Leonardi's carefully observed details of the anatomy of the common porcupine, drawn in the early seventeenth century.

From start to finish, this book is so unremittingly silly that it occurred to me that Evans set out to write a sort of post-modernism-inspired spoof, a demonstration that poppycock can get published and passed off as science. I expect that most medical scientists will dismiss the book out of hand. For those who want an informative, scholarly and comprehensive view, I recommend *The Placebo Effect: An Interdisciplinary Exploration*, edited by Anne Harrington (Harvard University Press, 1997).

The author does not tell us the intended audience for his book, but he supplies definitions of common medical terms and explanations of medical concepts, so I suspect that he is writing for a general audience. Many educated general readers will recognize his empty theorizing, circular reasoning and mis-statements for what they are. But there has always been a ready market for ideas that in a single stroke explain the ills that afflict us and reveal the pathway to health. In modern times these notions often come with the trappings of science and a whiff of scientific plausibility. Some will find in Evans a kindred spirit, and feel their hearts warmed by the knowledge that even something as seemingly elusive as the placebo response can be easily explained. ■

Walter A. Brown is in the Department of Psychiatry and Human Behaviour, Brown Medical School, and Tufts University School of Medicine, and is at 108 Driftwood Drive, Tiverton, Rhode Island 02878, USA.

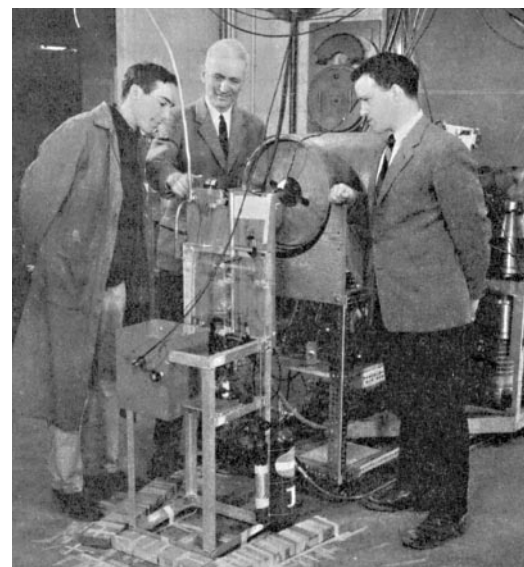
First among atoms

Hydrogen: The Essential Element
by John S. Rigden

Harvard University Press: 2002. 288 pp.
\$28, £19.50, £28

Daniel Kleppner

It would be hard to find a better icon for reductionism than the hydrogen atom. It has been said, as John Rigden recalls in *Hydrogen*: "To understand hydrogen is to understand all of nature." Here he provides abundant evidence to support this extravagant statement. The atom that played a seminal role in the creation of key theories by Niels Bohr, Werner Heisenberg, Erwin Schrödinger and Paul Dirac, not to mention in relativistic quantum electrodynamics, surely occupies a special place in the history of science. Hydrogen makes up most of the known matter in the Universe, and its radiation has told us a great deal about our galaxy and the cosmos. And hydrogen has played principal roles in magnetic resonance and atomic clocks. Such an atom demands to have its story told, and here



Hydrogen — the most abundant element in deep space, where it forms huge gas clouds (left) — is used in maser clocks, shown above with Daniel Kleppner (right) and Norman Ramsey (centre).

Rigden does an admirable job of telling it.

Rigden narrates the events surrounding discoveries that are generally glossed over in the classroom. He describes how a pythagorean obsession led Johann J. Balmer to discover his prescient empirical formula for hydrogen's wavelengths, and how the fine-structure constant emerged from a failed theory — Arnold Sommerfeld's attempt to amalgamate the Bohr model and relativity. This was one of the great failures in physics, and seems to contradict Einstein's dictum that God is subtle but not malicious.

Hydrogen features strongly in Rigden's account of events at Columbia University in New York during the golden years of Isidor I. Rabi from 1930 to 1950. Rabi attempted to measure the magnetic moment of the proton by improving the molecular-beam method he had brought to the United States from Otto Stern's laboratory in Hamburg. In 1937 this effort culminated in his invention of molecular-beam magnetic resonance, which opened up a new world for spectroscopy and inspired developments such as nuclear magnetic resonance.

It is difficult to overestimate the impact of this invention, both direct and indirect. An early experiment led by a young graduate student, Norman F. Ramsey, indicated that the deuteron (a deuterium nucleus) has a quadrupole moment. This first evidence for non-central nuclear forces constituted a great advance in nuclear physics.

Another experiment revealed that the hyperfine structure of hydrogen disagrees with the value predicted by the Dirac theory. Resolving this discrepancy inspired Julian Schwinger to formulate his version of quantum electrodynamics. Willis Lamb's

microwave measurement of the Lamb shift, made in a nearby laboratory, provided further inescapable evidence that the Dirac theory was deficient. And Rabi's magnetic-resonance methods contributed to the discovery of nuclear magnetic resonance by Edward M. Purcell and Felix Bloch, and the maser and laser by Charles H. Townes, Arthur L. Schawlow and others.

Rigden describes how hydrogen maser atomic clocks are used for deep-space navigation, global positioning and in time-keeping laboratories. He narrates the million-fold advance in spectroscopic precision by Theodor W. Hänsch in his monumental studies of hydrogen. This quest inspired Hänsch's invention of the frequency comb, a revolutionary technique for measuring optical frequencies; the impact of this invention is only starting to be felt. Another advance, too recent for this book, is the creation of antihydrogen atoms at CERN, the European laboratory for particle physics near Geneva. The story of hydrogen is by no means finished.

Hydrogen is aimed at both the public and students of physics. It does not strive for detail at the level of Abraham Pais' *Inward Bound* (Clarendon Press, 1986), nor does it indulge in speculation — for instance, about whether the creed of reductionism is likely to lose prominence in the coming century, causing the icon of the hydrogen atom to dwindle into a mere curiosity. But its friendly size, straightforward style, numerous illustrations and pithy aphorisms all make for engaging reading. ■

Daniel Kleppner is in the Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

Behind the blooms

What was your first experiment as a child?

Changing the colour of petal pigments using spirits of salts (hydrochloric acid, from my father's soldering kit) and caustic soda (sodium hydroxide, from my mother's soap-making cupboard).

Who has been the most important mentor in your career?

David Catcheside, a great molecular geneticist, a key figure in establishing genetics in Australia, and whose students understood the difference between a 'fact' and a 'model'.

What single scientific paper or talk changed your career path?

As an undergraduate I read Ed Lewis' paper on homeotic mutants of *Drosophila* (*American Zoologist* 3, 33–56; 1963). It took 25 years, but eventually we followed his pioneering path in plants.

What book has been most influential in your scientific career?

C. W. Wardlaw's *Organization and Evolution in Plants* and Darwin's *The Voyage of the Beagle*.

Is there a 'tyranny of reductionism' in how scientists are trained today?

Are scientists 'trained'? Yes, they are exposed to facts; they gain experience in setting up and interpreting experiments; and they see others in action. But whether or not they look above their bench or computer screen is mostly an inherent trait, and is very difficult to promote if the desire is absent.

What's the one thing about science that you wish the public understood better?

Science is based on the joint pillars of curiosity and evidence; and the outcomes are unpredictable.

If you could direct more government funding into one area of science, where would you put it?

Pure mathematics, the ultimate creative science.

Whose graduate student would you most like to have been (historical impossibility notwithstanding)?

Thomas Hunt Morgan, for the joy of being there as the wider principles of inheritance were being deduced.

What gives you the most job satisfaction now? What are your major frustrations?

Seeing graduate students come alive as they discover the joy and excitement of doing real research makes it all worthwhile. On the other hand, increasing pressure to do research in which the answer is apparently already known

and 'just ripe for an application that will make money for someone' is frustrating and requires delicate avoidance skills.

What literary character would you employ as a postdoc?

Elizabeth Bennet from Jane Austen's *Pride and Prejudice* would fit in beautifully, but she would need to learn how to use a Gilson Pipetman.

What's your favourite conference destination, and why?

Taos, New Mexico, because of its smallness, remoteness, social atmosphere and beautiful physical surroundings — delegates really do interact.

What was the worst/most memorable comment you ever received from a referee?

"The paper is spoiled ... by incomplete and sometimes sloppy molecular analysis." This has been very useful in pushing current students into tidying up their molecular results.

You have the audience in your hands, but some smart-alec asks you the killer question you have no idea how to answer. What's your favourite response?

When posed just such a question, Cyril Darlington (a British cytogeneticist) replied, "I thought that is what I had been trying to say all along", and then immediately looked round the audience for the next question.

What book is currently on your bedside table?

Andrew Robinson's *The Story of Writing*.

What music heads the playlist in your car or lab?

I enjoy silence most of the time (I must be a mutant).

Assuming the dead can be raised and/or time travel exists, who from the world outside science would you most like to have dinner with?

The famous Australian anthropologist Baldwin Spencer and his lay colleague Frank Gillen, who learned about aboriginal life while he was postmaster at Alice Springs.

Where and when would you most like to have lived or worked?

Antarctica in 2050, sufficiently distant in the future to see how it all turned out for Earth and the human race, and perhaps the only place remaining cool enough to live comfortably!

What do you most dislike about having research published?

The artificiality of having to make it into a 'story' (see Peter Medawar's essay *Is the scientific paper a fraud?*).



David Smyth

David Smyth is professor of genetics at Monash University in Melbourne, Australia. His group has identified new plant transcription factors that control flower development in *Arabidopsis*.

What overlooked or underrated discovery really changed the science in which you work?

The major role that RNA plays in regulating gene function was first uncovered in plants (from studying coloured patches in petunia petals). I suspect the key original role that green organisms played will be forgotten as such studies rapidly revolutionize molecular genetics.

What's the best piece of advice you've ever received?

"Don't spend your life doing with calcium what has already been done with magnesium." (Excellent advice from David Hayman, my undergraduate mentor.)

What would you have become, if not a scientist?

A historian. Historians help us to place our own times in perspective. For example, it was a revelation for me to spend some time in Israel and Jordan in 2000, and to learn about the 9,000 years of history of the Middle East that nurtured so much of present Western culture.

Name one extravagance you can now get away with because of your eminence.

Eminence, what eminence?

What music would you have played at your funeral?

Bach's *Concerto for Two Violins* (BWV 1043).

What's just around the corner?

Nuclear energy to replace the burning of fossil fuels that are endangering our climate by the relentless pumping of carbon dioxide into the atmosphere. We have fixed the ozone hole, now let's look after the greenhouse effect. ■

The key to signalling

Tetsuo Kushi, Eiji Nambara and Peter McCourt

Cells of multicellular organisms need to communicate with each other to regulate their development and organize growth and cell division. Hormones contribute to these processes by acting as messengers between cells, telling them what's happening elsewhere and how they should respond. The manipulation of hormones can have profound effects on society — for example, the birth-control pill and the high-yielding crops that have fed the world since the 'green revolution' both rely on the manipulation of specific hormone functions. On the flip side, the use of performance-enhancing steroids is the bane of modern athletic competitions.

Among the many types of hormone, one particular class — the water-insoluble or hydrophobic hormones, which include mammalian steroids and many plant-growth regulators — is particularly interesting because the chemical structures of its members are often conserved between plants and animals. This similarity is surprising because it is generally thought that the plant and animal kingdoms separated from each other in evolutionary terms long before either group became multicellular. As a consequence, multicellular growth and development in these two kingdoms are very different. So why are similar keys (chemical structures) used to unlock fundamentally different developmental locks?

The conservation of biochemical molecules between say, mice and trees, could mean that common biosynthetic pathways existed before the plant and animal kingdoms separated, and that evolution has simply used these ancient pathways to create similar-looking signalling chemicals. This seems to be true for steroid-like hormone synthesis, as plants and animals have many biosynthetic steps in common. Other hormones are derived from common precursors found in plants, such as the animal retinoids and

prostaglandins and the plant growth regulators abscisic acid and jasmonate. The use of common precursors certainly constrains the overall structure of hormones, but later synthetic steps are also often conserved to some degree, suggesting that the final structures have some inherent properties that make them good signalling molecules, regardless of the organism in which they function.

Part of the answer may lie with the origins of these structures. It has been speculated that the membranes of ancient, single-celled organisms were not the fatty-acid bilayers of today, but were more probably composed of terpenoid-based compounds. Perhaps steroids and retinoids, which are terpenoid derivatives, evolved from these early barrier membranes when organisms needed to know what was going on outside. If so, these hormones are some of the oldest types of signalling molecule. However, many of today's hydrophobic hormone structures require oxidation steps late in their biosynthesis. Molecular oxygen was not abundant in the atmosphere until after the development of photosynthetic organisms around 3.5 billion years ago, and so these final structures are unlikely to have evolved until after this time. Modern hydrophobic hormones may not therefore represent the original signalling hormones.

Every signal needs a receiver, and hormones in plants and animals are detected by protein receptors. In animals, steroid hormones interact with protein receptors within the cell that often bind directly to DNA to regulate gene expression. Despite the availability of completely sequenced genomes of both plants and animals, we cannot find anything that looks like an animal steroid or retinoid receptor in any plant genome. The one steroid receptor (brassinosteroid leucine-rich repeat receptor kinase) that has been identified in plants is completely unrelated to any of the known animal steroid receptors. So although hormone structures in plant and animals may have common origins, the receptors that recognize them probably evolved independently after the two kingdoms separated.

This raises some intriguing possibilities, because a signalling molecule needs something to signal to or else it would be discarded by evolution. Perhaps, then, hydrophobic hormones recognized some other ancient receptors. The recent reports of organic molecules binding directly to endogenous RNA sequences, turning them into 'riboswitches', indicates that RNA has the necessary specificity to be a receptor. This capacity could have predated the existence of complex proteins. The fact that several screens for genes involved in the response to abscisic acid have identified genes involved in RNA processing may hint

Hormone evolution

The similarity of hydrophobic hormones in plants and animals suggests that once you make a good key, with occasional filing it can be used in many different developmental locks.

at a link between this retinoid-like hormone and RNA in plants. If RNA hormone receptors did precede their modern protein counterparts, they would probably have regulated translation, whereas the first protein receptors probably bound directly to DNA, to regulate transcription. The regulatory DNA sequences or promoters that normally control transcription are usually a composite of binding sites for different proteins, thereby allowing a gene to respond to various inputs. Such input flexibility may have been more difficult to evolve at the level of translation.

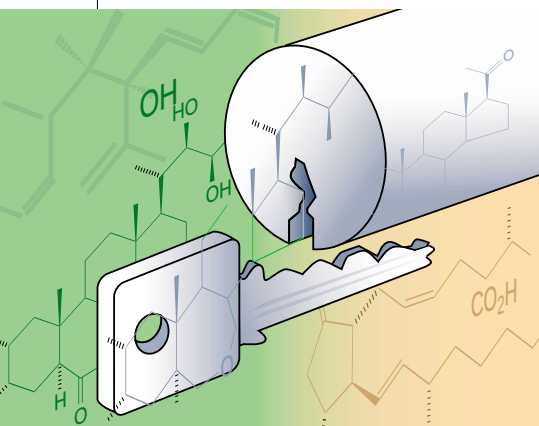
Alternatively, if protein hormone receptors came into existence at the same time as the hormones they recognize, then coevolution of a receptor and its complementary hormone would be expected. Interestingly, although hydrophobic hormones are chemically diverse, their sizes are quite similar. Maybe the hormone-binding pocket of the first receptor constrained the overall size of any new chemical that could find use as a signalling molecule. This idea has been used to explain why one superfamily of related nuclear receptors seems to perceive all of the different hydrophobic hormones in animals. By analogy, perhaps the missing receptors for many of the plant hydrophobic hormones, such as abscisic acid and gibberellins, are encoded in the large family of genes that are related to the brassinosteroid leucine-rich-repeat receptor kinase.

Just as the filing of a hormone key may allow the opening of vastly different receptor locks, it may also be that a good lock can be used by many different keys.

Tetsuo Kushi and Eiji Nambara are at the Laboratory for Reproductive Growth Regulation, Plant Science Centre, RIKEN 2-1 Hirosawa, Wako, Saitama, Japan.
Peter McCourt is in the Department of Botany, University of Toronto, Toronto M5S 3B2, Canada.

FURTHER READING

- Bogan, A. A., Cohen, F. E. & Scanlan, T. S. *Nature Struct. Biol.* **5**, 679–681 (1998).
- McCourt, P. *Mol. Cell* **8**, 1157–1158 (2001).
- Mussig, C. & Altmann, T. *Trends Endocrinol. Metabol.* **12**, 398–402 (2001).
- Ourisson, G. & Nakatani, Y. *Chem. Biol.* **1**, 11–23 (1994).
- Szostak, J. W. *Nature* **419**, 890–891 (2002).



Flowers' wings, fruitflies' petals

Claude Desplan and Thomas Lecuit

The technique of clonal analysis allows dividing cells in a developing organ to be marked and tracked. The aim of such studies, in plants as well as animals, is to understand how tissues acquire their form and size.

Over the past decade, molecular genetics has started to deliver answers to the age-old question of how an organism acquires its final design. The work concerned has shown that morphogen molecules 'pattern' tissues¹, and so has provided a clear view of how cell identity is spatially and temporally defined in a growing organism. But this patterning has remained an abstract concept, based simply on how different domains of gene expression are established immediately downstream of a particular morphogen. How these molecular events lead to the overall form of the tissue, often several days later, has remained a black box.

Writing on page 161 of this issue², Rolland-Lagan, Bangham and Coen offer a fresh view of the problem. The authors analysed a simple plant tissue, the petal lobe of *Antirrhinum* (snapdragon), to describe how a tissue acquires its form. The petal is particularly amenable to this study because of its simple organization, and because cell movements, which are involved in the acquisition of form in animals, are clearly extremely limited in plants. Rolland-Lagan *et al.* provide a simple, quantitative, dynamic model in which just a few parameters are sufficient to explain the final shape of the tissue.

The authors used classical clonal analysis in their experiments. They generated and followed the fate of small patches of marked tissue distinguished, for instance, by colour, each being the progeny of an individual cell (Fig. 1a). By comparing average characteristics of clones induced at various times for each region of the petal, they could compute growth rate (that is, the time between two cell divisions), the extent of anisotropy of growth (how much the descendants of dividing cells accumulate preferentially along a given axis) and the main direction of growth.

Rolland-Lagan *et al.* measured the average values and spatiotemporal fluctuations of these parameters. They then assessed the significance of the experimental measurements using computer simulations. The first notable conclusion is that the estimated parameters (growth rate, anisotropy and direction of growth) are sufficient to generate the correct changes in petal shape. Remarkably, the observed localized differences in cell division rates, anisotropy or direction are not essential for generating petal asymmetry, as the simulation works well with fixed average values for these three

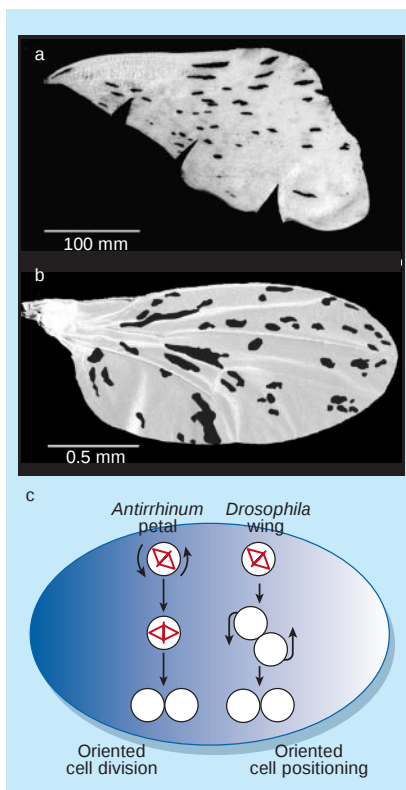


Figure 1 Clonal analysis and organ shape.

a, *Antirrhinum* petal, for which Rolland-Lagan *et al.*¹ inferred patterns of cell divisions from the shape of marked clones of cells. b, The *Drosophila* wing, which has been subject to similar analyses. Both petals and wings display a marked asymmetry in clone growth, possibly in response to unknown long-range signals. c, In the *Antirrhinum* petal, such long-range signals could orient cell division, whereas in the *Drosophila* wing the plane of division is random but cell positioning is oriented. In each case, the result is asymmetric clone growth along the main tissue axis. a is a light-microscope image and b is a scanning electron micrograph; courtesy of A.-G. Rolland-Lagan and K. Lee, respectively.

types of parameter. The authors also suggest that long-range signals orient cell divisions. Overall, these results appear to be amazingly simple and make it possible to identify the genetic components that control the three parameters. This could lead to a generalized model for other tissues, not only in plants but also in animals, even if more parameters are necessary.

Since the early 1970s, the wing of the fruitfly *Drosophila* has also been subjected to clonal analysis to address the problem of how tissue shape is controlled. Garcia-Bellido and co-workers, among others, have studied how the shape and size of the fly wing is created by cells dividing in the precursor of the developing wing, the imaginal disc. The comparison between this system and the flower petal is striking (Fig. 1a, b). Although they did not reach the same level of quantitative analysis, some of the conclusions of these earlier investigations^{3,4} are evidently similar to those of the latest studies¹, and some are different.

The main similarity is that cell divisions occur throughout the tissue rather than as a wave or in a growth zone. Clone proliferation is also anisotropic, although the direction of asymmetric growth changes at different stages in the developing fly wing. Clones generally grow along the main tissue axes (Fig. 1a, b), reflecting the overall tissue shape. Finally, both systems seem to rely on long-range signals (Fig. 1c). For the flower petal, the long-range signal could specify the main direction of cell growth and division. In the fly wing, morphogens such as the Decapentaplegic or Wingless proteins could also direct the ordered positioning of cells along the morphogenetic gradients, although they do not affect the plane of cell divisions, as their orientation seems to be random^{4,5}. There are, however, variations on this theme and Garcia-Bellido's 'entelechia' model⁵ calls for local monitoring of positional values that control cell division and cell positioning, independent of a long-range gradient. Here, an interesting parallel can be drawn with the establishment of tissue polarity⁶, in which long-range signals orient planar cell polarity acquired by self-propagating, local cell-cell interactions.

What else regulates tissue shape and size? In particular, what is the contribution of cell number and cell size? This question has been addressed in both plants and animals, where mutations in the *cdc2* gene limit cell division but organ size is maintained through a compensation mechanism that increases cell size^{4,7,8}. Moreover, several mutations that severely distort the relative size of individual clones fail to alter the shape and size of the organ in the fly (for instance, mutations in *Minute*⁹). These findings suggest that global dimensions are also measured in an organ, or

even at the level of the whole organism. How this is done is a fundamental question that remains unanswered.

Plants and animals might have invented multicellularity independently. Yet despite the obvious differences at the molecular level, there are strong similarities in their strategies for controlling organ shape and size. At a higher level again, however, further controls exist that are fundamentally different between plants and animals — those that determine the relationship of organ size to organism size (allometry). As a plant grows larger, its mature flowers or leaves stay the same size. In animals, all organs grow proportionally. No doubt future quantitative studies on different systems, as offered by Rolland-Lagan *et al.* for the petal, will pave the way for a more complete understanding of this fascinating phenomenon. ■

Claude Desplan is in the Department of Biology, New York University, 1009 Main Building, 100 Washington Square East, New York, New York 10003-6688, USA.
e-mail: claudedesplan@nyu.edu

Thomas Lecuit is at the IBDM/LGPD, Campus de Luminy, Case 907, 13288 Marseille Cedex 09, France.
e-mail: lecuit@ibdm.univ-mrs.fr

1. Lawrence, P. A. & Struhl, G. *Cell* **85**, 951–961 (1996).
2. Rolland-Lagan, A.-G., Bangham, J. A. & Coen, E. *Nature* **422**, 161–163 (2003).
3. Resino, J., Salama-Cohen, P. & Garcia-Bellido, A. *Proc. Natl Acad. Sci. USA* **99**, 7502–7507 (2002).
4. Day, S. J. & Lawrence, P. A. *Development* **127**, 2977–2987 (2000).
5. Garcia-Bellido, A. C. & Garcia-Bellido, A. *Int. J. Dev. Biol.* **42**, 353–362 (1998).
6. Ma, D., Yang, C. H., McNeill, H., Simon, M. A. & Axelrod, J. D. *Nature* **421**, 543–547 (2003).
7. Hemerly, A. *et al.* *EMBO J.* **14**, 3925–3936 (1995).
8. Weigmann, K., Cohen, S. M. & Lehner, C. F. *Development* **124**, 3555–3563 (1997).
9. Simpson, P. & Morata, G. *Dev. Biol.* **85**, 299–308 (1981).

Astronomy

Atmosphere out of that world

David Charbonneau

A planet orbiting very close to a Sun-like star is apparently enveloped by an extended atmosphere of hydrogen atoms, and may be losing mass because of the intense radiation from the parent star.

Of the one hundred or so planets known to orbit nearby Sun-like stars, one is unique. Its name is HD209458b, and the plane of its orbit is fortuitously tilted so that, once every revolution, observers on Earth see the planet in eclipse — in ‘transit’ — in front of its star. Although the planet is similar to Jupiter in mass and size, it is a hundred times closer to its star than Jupiter is to the Sun. As a result, the planet’s gravity must battle the intense heat and stellar radiation to retain its atmosphere.

On page 143 of this issue, Alfred Vidal-Madjar and colleagues¹ present evidence that the planet may be losing this fight. Based on observations by the Hubble Space Telescope, they infer that an extended envelope of hydrogen surrounds the planet and suggest that it is a ‘cometary tail’ of material being evaporated from the object. As a result, this planet may slowly be losing mass over time. The implication is that planets initially located even closer to their stars would not survive long, a conclusion that agrees with the observed paucity of extrasolar planets in such orbits.

Following the first detection² of a distant planet orbiting a Sun-like star in 1995, plans were soon made for the spectroscopic investigation of possible planetary atmospheres. These studies stemmed from the desire to obtain direct information on the composition and local conditions of such planets. Researchers realized that the data would yield insights into the formation and

dynamical evolution processes that had shaped these bodies, and might reveal whether such events had been similar to, or different from, those that had occurred in our own Solar System. But the direct approach — first isolating the light from a distant planet, and then performing spectroscopy on it — still remains beyond our technological reach. The stumbling-blocks are the very small angular separation of the

planet and parent star (as observed from Earth), coupled with the extremely high contrast ratio of the pair. For comparison, consider that although Jupiter is the brightest planet of the Solar System, it shines in reflected sunlight with an intensity that is only a few parts per billion that of the Sun.

With the discovery^{3–5} of the first extra-solar planet to transit its parent star, several researchers^{6–8} quickly identified a golden opportunity to sidestep the technological limitations described above. During transit, some of the light from the parent star filters through the outer reaches of the planetary atmosphere, and has impressed upon it the additional spectroscopic absorption features of the atmospheric constituents. By taking the ratio of stellar spectra observed in and out of eclipse, the stellar features are removed, and those of the planet are isolated. With this technique, the glare of the central star is no longer a hindrance, but rather a probe of the planetary atmosphere.

This method had its first success in 2001, when colleagues and I detected⁹ sodium in the atmosphere of HD209458b. But although sodium is spectroscopically very active, it is only a trace component of the atmosphere. Hydrogen, on the other hand, is the majority constituent of this gas-giant planet, and thus has much greater diagnostic potential. With this in mind, Vidal-Madjar *et al.*¹ observed three separate passages of HD209458b in front of its parent star. During each event, they measured a 15% dimming in the starlight (Fig. 1) emitted at the wavelength corresponding to a fundamental transition of the hydrogen atom. The large amplitude of this signal is somewhat of a surprise. The planet’s radius is roughly 1.35 times that of Jupiter, so that the planet blocks only 1.5% of the area of the star — indeed,

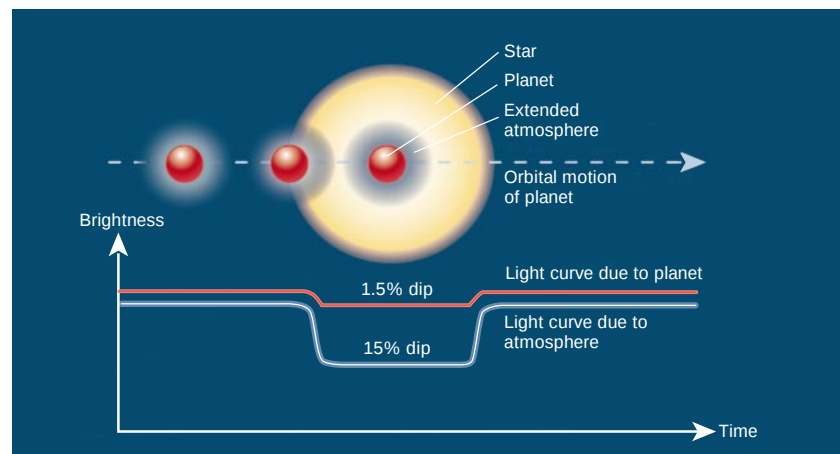


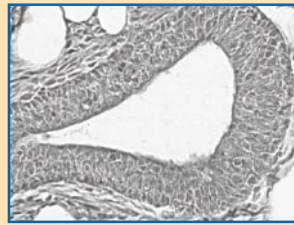
Figure 1 The observations from which Vidal-Madjar *et al.*¹ conclude that the planet HD209458b has an extended atmosphere. Passage of the planet in front of its parent star blocks 1.5% of the starlight when observed at most wavelengths, producing the small characteristic dip in the light curve (red line). Vidal-Madjar *et al.* observed a much deeper transit depth of 15% (blue line) at the wavelength of a transition of atomic hydrogen, from which they deduce that an atmosphere of hydrogen envelops the planet. It appears to extend beyond the gravitational reach of the planet, implying that some of the hydrogen atoms are escaping.

Developmental biology

Guidance molecule goes global

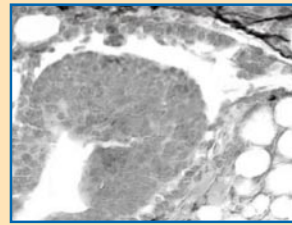
The guidance molecule netrin-1 is famous for its role in telling axons — the long extensions sent out by nerve cells — where they should and shouldn't go in the nervous system. Reporting in *Developmental Cell* (4, 371–382; 2003), Karpagam Srinivasan *et al.* extend the repertoire of netrin-1 activities to non-neuronal tissues, showing that it also acts to keep cells stuck together during the development of mammary glands.

It was already known that netrins are secreted by many cells outside the nervous system, but until now no one had really worked out what they were doing. Srinivasan *et al.* looked more closely at where netrin-1 occurs in the



developing mammary glands of mice, and found that it surrounds the cap cells — the single layer of cells that caps the developing gland, or bud — in a pattern that is complementary to one of its receptors, neogenin.

Extrapolating from netrin-1's function in the nervous system, it might be predicted that it provides a positional cue to guide moving cells



within the mammary gland.

Surprisingly, the authors found that, instead, netrin-1 prevents cap-cell movement. Loss of either netrin-1 or neogenin disrupted adhesion between the cap-cell layer and adjacent cells, and resulted in cap cells moving into regions where they would not normally go, as seen in these images of normal (left) and netrin-deficient (right) buds.

Furthermore, addition of netrin-1 to isolated neogenin-producing cells caused them to aggregate. So it seems that netrin-1 may be required in the developing mammary gland simply to make sure that cells stick together.

The authors are now investigating the long-term consequences of loss of netrin-1 for mammary-gland development, and in particular the possibility that these tissue disruptions increase susceptibility to cancer. At a more basic level, how does the binding of netrin-1 to neogenin immobilize cells? Studies of netrin-1 in the nervous system provide some hints, but there may be more surprises in store. **Alison Schuldt**

this small transit depth (Fig. 1) has been observed repeatedly^{3,4}. The amplitude of the transit depth reported by Vidal-Madjar *et al.* thus corresponds to an extended atmosphere that has a radius 4.3 times that of Jupiter. At this distance, the gravity of the planet is no longer sufficient to retain the hydrogen atoms, and so some fraction trickles away. The calculated minimum escape rate for this process would reduce the planetary mass by only a negligible amount (0.1%) if held constant over its roughly 5-billion-year age. However, the data indicate a higher escape rate, enough to trim the planetary mass significantly.

Vidal-Madjar *et al.* conclude by pointing out that this planet lies exceptionally close to its parent star (completing a full orbit once every 3.5 days). The authors speculate that, at smaller orbital separations, the rate of mass loss is correspondingly higher, and planets that initially reside in such orbits evaporate. Eight extrasolar planets have similar orbital periods (between 3 and 4 days), and only recently has a planet been found¹⁰ that has a shorter period (of 1.2 days). Seared to a temperature of 1,900 K, this new-found planet is in the hot seat — its existence provides an upper limit on the evaporation rate.

Reservations about Vidal-Madjar and colleagues' findings centre on possible contaminating sources of emission that vary with time at the wavelength concerned. Sunlight scattered by the outermost layers of the Earth's atmosphere is large and variable as viewed from the orbiting Hubble Space Telescope. And the emission of the parent star itself may vary in time or across the stellar surface. The authors present good evidence

that these effects are excluded, but their conclusions will be all the stronger with confirmatory data.

Modelling the outer reaches of the planetary atmosphere in this high-energy environment, taking these new observations into account, should prove rewarding for both planetary scientists and astronomers, and there is always the promise of more data to come. The transiting configuration of HD209458b will ensure that it remains the centre of attention in the immediate future, and a keystone in our understanding of planets outside the Solar System. ■

David Charbonneau is at the California Institute of

Technology, 105-24, 1200 East California Boulevard, Pasadena, California 91125, USA.
e-mail: dc@caltech.edu

1. Vidal-Madjar, A. *et al.* *Nature* **422**, 143–146 (2003).
2. Mayor, M. & Queloz, D. *Nature* **378**, 355–359 (1995).
3. Henry, G. W., Marcy, G. W., Butler, R. P. & Vogt, S. S. *Astrophys. J.* **529**, L41–L44 (2000).
4. Charbonneau, D., Brown, T. M., Latham, D. W. & Mayor, M. *Astrophys. J.* **529**, L45–L48 (2000).
5. Mazeh, T. *et al.* *Astrophys. J.* **532**, L55–L58 (2000).
6. Seager, S. & Sasselov, D. D. *Astrophys. J.* **537**, 916–921 (2000).
7. Brown, T. M. *Astrophys. J.* **553**, 1006–1026 (2001).
8. Hubbard, W. B. *et al.* *Astrophys. J.* **560**, 413–419 (2001).
9. Charbonneau, D., Brown, T. M., Noyes, R. W. & Gilliland, R. L. *Astrophys. J.* **568**, 377–384 (2002).
10. Konacki, M., Torres, G., Jha, S. & Sasselov, D. D. *Nature* **421**, 507–509 (2003).

Behavioural science

Fair's fair

Truman Bewley

A basic theory of behaviour holds that people act only in their own best interests. But more complex motives are apparent in an experimental study that shows that generosity is diminished by the unfairness of others.

The rationality hypothesis brings order to much of the thinking in social sciences, and especially in economics. According to this hypothesis, people act solely to advance their own interests, interpreted in the most selfish way. But everyday experience indicates that this is not entirely true: people are willing to make sacrifices to reciprocate favours or to take revenge. People tip waiters even though they will never see them again, and insults can lead to dangerous fights. Experimental evidence that supports these common-sense observations is accu-

mulating. On page 137 of this issue, Fehr and Rockenbach¹ describe experiments in which positive and negative forms of reciprocation — rewards and revenge — are potentially in conflict. They find that the threat of punishment for not rewarding a favour adequately can diminish the actual reward. Such insight into the motives that govern generosity, and our notions of fairness, is vital in the search for realistic principles that improve on the rationality hypothesis.

Applied to game theory, the rationality hypothesis gives rise to the concept of

'subgame perfection' — the assumption that people pursue their own interests from each point in a game onwards. In particular, players will not avenge or reward the actions of another player if doing so hurts their own interests at that time. People will never reciprocate kindness unless doing so brings further advantages to themselves. Nor will rational people avenge wrongs if the revenge will be costly to themselves. A great deal of evidence supports subgame perfection and the rationality hypothesis. For instance, in the workplace an employee's performance can be improved by offering financial incentives — or by disciplinary action.

But there is striking evidence against subgame perfection, in the form of experimental work on ultimatum games². In these, one player (the leader) makes an offer to another (the follower) on how to split a sum of money. If the follower accepts, the split is made according to the leader's offer. But if the follower refuses the offer, neither player gets anything. According to the principle of subgame perfection, the leader should get almost the whole amount, because the follower gains nothing by refusing any offer that gives him something. In practice, however, small offers are almost always refused.

Fehr and Rockenbach's experiments¹ are more subtle. The subjects of their study (a sample of more than 200 students at the University of Bonn) played a game involving two players, an investor and a trustee, each of whom was given an equal sum of money. The investor decided on an amount of money to give to the trustee and specified the amount that he wanted the trustee to return. The experimenter tripled the amount offered by the investor and passed it on to the trustee. The trustee then chose how much to return to the investor. In a second version of the game, the investor, when making the gift to the trustee, could commit to imposing a fine of a fixed size on the trustee if he returned less than the amount requested. Each investor and trustee interacted only once (they were recruited for the experiment on the spot in the university canteen), so that no player could reward or punish a partner's behaviour in future rounds of play.

On average, trustees reciprocated investors' generosity by making payments that increased with the size of investors' transfers. Trustees were least generous when the fine was imposed, more generous when there was no possibility of a fine, and most generous when the investor could impose a fine but chose not to do so. Fehr and Rockenbach's close examination of their evidence indicates that perceptions of fairness influenced the trustees' negative reactions to imposition of the fine. The earnings of the trustee and investor are equalized if the trustee returns two-thirds of the investor's transfer. Trustees apparently

made this calculation, for the imposition of a fine had less effect on what trustees returned when investors requested that the trustee return less than two-thirds of their transfer than when they requested that more than two-thirds be given back.

An explanation of the authors' findings is obvious — people are insulted and angered by threats that constrain their actions. Most of us want the freedom to choose. Nevertheless, the fairness that the authors emphasize may be crucial. Evidence from many sources indicates its importance³.

Planetary science

The core of planet formation

Bill Minarik

The rocky bodies from which the Earth formed may have already separated into a metal core and silicate shell. Innovative experiments exploring the behaviour of molten metal trapped between silicate grains suggest how.

Roughly speaking, the Earth is a metallic core surrounded by a silicate shell. Understanding the mechanisms that caused this separation, or differentiation, is one of the outstanding questions of Solar System science. Most of the Earth's volume is inaccessible to researchers, so information about its core must be gleaned indirectly. It comes from four main sources: remote sensing techniques using, for example, gravity and seismic waves; the study of core material from other Solar System bodies found on Earth as meteorites; inferences from the geochemistry of rocks formed from magmas that originated deep below the Earth's surface; and laboratory experiments at high pressures and temperatures, to simulate conditions approaching those at the core. This last is the approach taken by Takashi Yoshino and colleagues¹, who, on page 154 of this issue, present new constraints on the mechanism and timing of core formation.

Core formation probably occurred early in the history of the Solar System. Evidence for this comes, for example, from the decay of short-lived isotopes: decay of the hafnium isotope ¹⁸²Hf to tungsten (¹⁸²W) constrains the timing of core formation to the first 30 million years of the Solar System for all four bodies for which we have samples — Earth, the Moon, Mars and the asteroid Vesta^{2,3}. Earth formed from a nebula of dust and gas, as material clumped together to form kilometre-sized planetesimals, which then rapidly accreted into larger bodies (Fig. 1), ranging up to thousands of kilometres in diameter. The end point of the accretion process involved energetic collisions of large planetesimals. The ejected material from one such collision re-accreted in the proto-Earth's orbit to form the Moon.

Chemical data require that both the

People do seem to have in mind norms of behaviour for themselves and others, and try to enforce them. This vital topic of human motivation certainly deserves more experimental exploration.

Truman Bewley is at the Cowles Foundation and Department of Economics, Yale University, New Haven, Connecticut 06520, USA.

e-mail: truman.bewley@yale.edu

1. Fehr, E. & Rockenbach, B. *Nature* **422**, 137–140 (2003).
2. Güth, W., Schmittberger, R. & Schwarze, B. *J. Econ. Behav. Org.* **3**, 367–388 (1982).
3. Moorman, R., Niehoff, B. & Organ, D. *Employee Responsibilities and Rights Journal* **6**, 209–225 (1993).

proto-Earth and the object with which it collided had already formed metallic cores. But data from meteorites are less clear. Samples of undifferentiated materials (such as from the chondritic meteorites) suggest that some planetesimals reached sizes of tens to hundreds of kilometres without substantial melting (and hence without separation); but other samples (from iron meteorites)

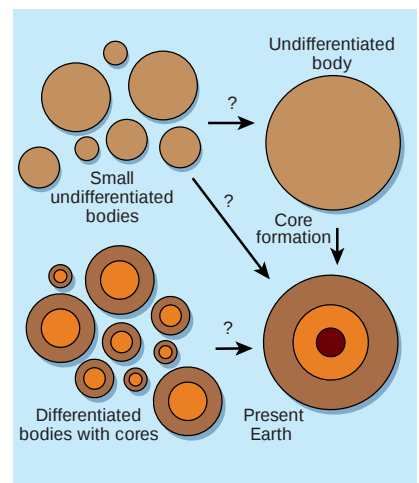


Figure 1 Timing of core formation. The Earth formed through accretion, absorbing planetesimals (lumps of rock and ice) through collisions. Did the Earth accrete undifferentiated material that then separated into shell and core — in which case, did the planet reach its present mass before differentiating, or was it a more gradual process? Alternatively, core formation might have happened rapidly inside growing planetesimals, so that the Earth's core is a combination of these previously formed cores. Isotopic evidence supports the latter model, and now Yoshino *et al.*¹ demonstrate a mechanism for the physical process.

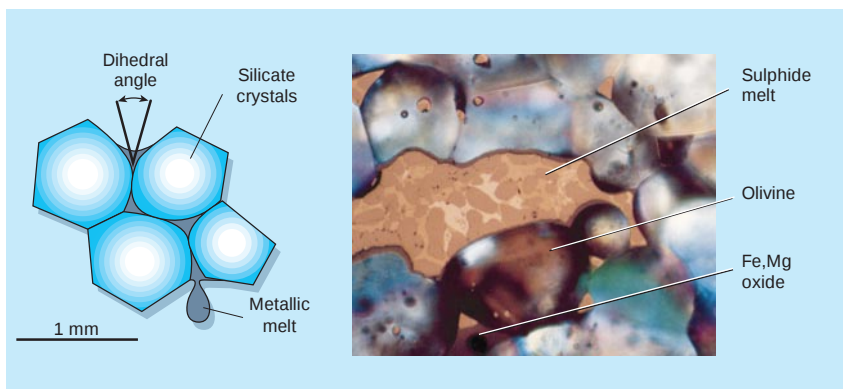


Figure 2 Degrees of separation. Metal alloys with a lower melting point than silicate minerals form a melt between the grain boundaries of silicate crystals. The dihedral angle (left) is a measure of how likely it is that pockets of melt will connect and separate from the silicate matrix. The photomicrograph (right) shows sulphide melt trapped in a matrix of Fe,Mg oxide and olivine (the silicate that makes up most of the Earth's upper mantle). The dihedral angle is near 90° and, if the melt makes up a sufficiently large fraction of the sample volume, the melt pockets start to interconnect. The field of view of the photomicrograph is about $500\text{ }\mu\text{m}$.

record differentiation into mantle and core.

The melting points of alloys of iron, nickel, sulphur and oxygen ($900\text{--}1,000^\circ\text{C}$) are lower than those of silicate minerals ($1,100\text{--}1,400^\circ\text{C}$), so when a cold planetary body is heated a metallic melt forms but the silicate and oxide minerals remain solid. Whether this metallic melt can segregate from the silicate matrix — and eventually form a metallic planetary core — depends on its microstructure, which is in large part determined by the interfacial energies of melt and minerals. These energies can be parametrized in terms of the dihedral angles made by the silicate grain boundaries around the pockets of metallic melt (Fig. 2). Systems with large melt–solid interfacial energies (relative to solid–solid grain-boundary energies) have large dihedral angles, resulting in isolated melt pockets surrounded by melt-free grain boundaries. In such systems, the melt is unlikely to separate from the matrix.

Constraints on this core-forming process mostly come from experiments, as meteorites have lost much of their crucial textural information by the time they reach Earth. Olivine — $(\text{Fe,Mg})_2\text{SiO}_4$ — makes up much of the Earth's upper mantle, and presumably made up much of the silicate fraction of accreting planetesimals. Metallic melts within an olivine matrix have large dihedral angles. This implies that, in a matrix that does not deform, core-forming melts would be trapped at grain boundaries. So for the melt to migrate towards the core, higher temperatures are needed to melt a matrix made of olivine and other silicate minerals. This process would produce a different trace-element chemistry in the melt than if the molten metal had percolated through a solid silicate matrix to the core.

Characterizing the permeability of a system using dihedral angles poses several

difficulties. First, there are theoretical assumptions. For instance, a single dihedral-angle value for a sample implies that the interfacial energies are isotropic (which is not generally true in a system containing many different, anisotropic minerals); we must assume that the microstructure is dominated by surface reaction kinetics, not by mechanical deformation. In experiments, the system geometry must remain stable during the transition from high temperature and pressure to ambient conditions. Finally, with increasing melt fraction, even a system of isolated melt pockets will start to interconnect, but the 'percolation threshold' — the minimum melt fraction required for interconnection — has been modelled numerically only for simple systems⁴.

To overcome these limitations, Yoshino *et al.*⁴ have performed the first high-temperature, high-pressure experiments that determine the interconnectivity of metallic melt in an olivine matrix through observed changes in electrical conductivity. Conductivity was measured after the samples were brought to a pressure of 3 gigapascals and a temperature of $1,200$ or $1,300^\circ\text{C}$, corresponding to a depth of about 100 km in the present Earth or of about 700 km in the Moon. The authors find that about 6% by volume of iron–sulphur melt is sufficient to form highly conductive pathways, representing interconnected (and therefore permeable) melt channels in the olivine matrix. Interestingly, although the formation, in systems with high dihedral angles, of a heterogeneous distribution of larger melt pockets surrounded by many melt-free grain boundaries has been predicted theoretically, and shown experimentally for other systems, Yoshino *et al.* saw no evidence for this in their experiments. This discrepancy is unresolved, and further experiments are needed.

If these high electrical conductivities



100 YEARS AGO

In Campbell Island, south of New Zealand, the breeze-fly (*Helophilus campbellicus*), one of the Syrphidae, so closely resembles a blow-fly (*Calliphora eudypti*) that when, in 1901, I captured a specimen of the first, which is rare, I thought it was the blow-fly, which is common... Now in any other locality this resemblance could be put down to mimicry. The blow-fly is common and offensive. The breeze-fly is rare and feeds on flowers. Everything favours this explanation except that in Campbell Island there are no insect-eating birds and no lizards, and consequently mimicry would be useless.

F. W. Hutton

Accidental resemblances between insects are to be expected... With regard to Captain Hutton's special instance, however, there appear to be certain points which require consideration before accepting the conclusion that the resemblance is merely a coincidence:— (1) The possible coexistence of the two species in other localities where the resemblance has a meaning; (2) the possible change of conditions in the struggle for life in the locality itself; (3) our possibly imperfect knowledge of the struggle which is waged there now.

B. Poulton

From *Nature* 12 March 1903.

50 YEARS AGO

The Nutrition Society held a symposium on 'The Role of Vitamins in Metabolic Processes' in the Biochemistry Department of the University of Sheffield on December 20, 1952... In the final paper, Dr. L. J. Harris (Dunn Nutritional Laboratory, Cambridge) examined the literature which, despite its bulk, still leaves us without evidence for a specific metabolic role of vitamin C. Among older hypotheses was one associating ascorbic acid with collagen formation, and also the more general theory that certain formative cells, such as osteoblasts, show reduced activity; in these cells the vitamin has been found to be concentrated in the Golgi apparatus... The great majority of animals can synthesize vitamin C, but, although glucose is known to be its precursor, it is still not known whether synthesis is a property of all cells of the body, or whether certain tissues are specialized for this purpose; similarly, the function of the vitamin in plants is still obscure.

From *Nature* 14 March 1953.

correspond directly to the permeability of the olivine matrix to metallic-melt percolation, the separation of a core from silicate mantle could happen very quickly in a planetesimal. The high temperatures required could result from the heat produced by the decay of short-lived isotopes present at the birth of the Solar System. More complete calculations of the thermal evolution of growing planetesimals (which include, for example, latent heat of melting, release of gravitational potential energy and impact kinetic energy) point to many sources of heat in the early Solar System that probably led to core formation and magma oceans in many growing planetesimals⁵.

But heating in a static environment may not be the whole answer. Deforming systems can have higher permeabilities than static systems, and impact-induced melting or differential stress may connect isolated melt pockets and produce pools of metal⁶ that may then sink through unmelted material. Each of these processes will tend to shorten the interval between accretion and core formation, so core formation should be ubiquitous once an accreting rocky planetesimal reaches a radius of 50–100 km. But then the existence of large bodies that do not seem to have differentiated (including some large asteroids such as Ceres, and Jupiter's moon

Callisto) is puzzling: is there some mechanism that prevented these bodies heating sufficiently to produce a core?

Building on the success of Yoshino *et al.*¹, future experiments may be able to determine melt connectivity through conductivity measured *in situ* and monitor dynamically evolving microstructure, such as during deformation or reactions. Synchrotron X-ray microtomography⁷ is another promising technique, which enables three-dimensional imaging with resolution approaching 1 μm^3 . These experimental advances will help us to understand the processes that have shaped the Solar System. ■

Bill Minarik is in the Department of Geology, University of Maryland, College Park, Maryland 20742, and at the Carnegie Institution of Washington, Washington, DC 20015, USA. e-mail: minarik@geol.umd.edu

1. Yoshino, T., Walter, M. J. & Katsura, T. *Nature* **422**, 154–157 (2003).
2. Yin, Q. *et al.* *Nature* **418**, 949–952 (2002).
3. Kleine, T., Münker, C., Mezger, K. & Palme, H. *Nature* **418**, 952–955 (2002).
4. von Bagen, N. & Waff, H. S. *J. Geophys. Res.* **91**, 9261–9276 (1986).
5. Ghosh, A. & McSweeney, H. Y. *Icarus* **134**, 187–206 (1998).
6. Bruhn, D., Growbner, N. & Kohlstedt, D. L. *Nature* **403**, 883–886 (2000).
7. Roberts, J. J., Tyburczy, J. A., Locke, D., Minarik, W. & Kinney, J. *Eos* **83**, F400 (2002).

Evolutionary biology

Teeth as tools

Anne Weil

What determines the shapes of mammalian teeth? When tools are designed to cut to the meat of the question, form follows function rather than developmental or evolutionary constraints.

Very different groups of mammals have teeth of similar shapes. One obvious explanation for this is that the greatest efficiency in chewing similar foods is strongly favoured by natural selection. But other reasons could include the constraints imposed in the process of development, or the historical limitations imposed within mammal lineages, and each of these factors might act against or in concert with the demands of optimizing function.

Writing in the *Biological Journal of the Linnean Society* (78, 173–191; 2003), Alistair R. Evans and Gordon D. Sanson describe how they have taken an engineer's approach to this question. They have carried out a computer-modelling exercise, designing tools to cut tough substances, and find that the most efficient tools closely resemble the molars of carnivorous and insectivorous mammals. They conclude that in many cases developmental and evolutionary factors have not strongly influenced molar shape, and that function is indeed the primary determinant.

Most mammals use differentiated cheek

teeth for chewing, to divide food into small pieces that can be swallowed easily and digested efficiently. Mammalian teeth are replaced at most once in an individual's lifetime, so exact positioning of them is possible, allowing the cutting edges and points of upper and lower teeth to meet in a precise way. The hands, tongue and facial muscles are variously used to position food between the teeth. Tough (as opposed to brittle or soft) foods are divided by an initial puncture (or punctures), which is then extended into a longer cut. Chewing tough foods can thus be envisaged as a mechanical task in which the teeth act as simple, edged tools.

Evans and Sanson considered six functional factors used by engineers in tool design: sharpness of points; sharpness of blades; the angle between the blade and the substance cut; the angle between the blade and a line perpendicular to the cut; the entrapment of substance between blades; and the movement of substance away from the blade that prevents the implement from becoming clogged up. They considered these

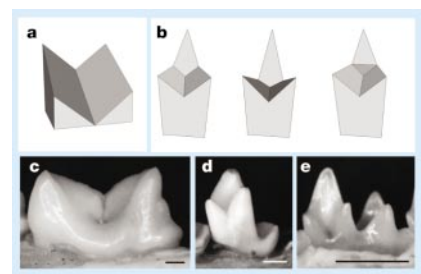


Figure 1 Efficient model cutting tools, and some similarly shaped mammalian cheek teeth. The single-bladed (a) and double-bladed (b) models optimize eight efficiency factors considered by Evans and Sanson. On that evidence, real mammalian molars (c, single-bladed; d, e, double-bladed) may approach a functional ideal. Scale bars, 1 mm.

factors for tools with single blades (which are rectangular in horizontal cross-section, and can resemble a chisel, scissors or a guillotine), as well as for tools with two blades that meet at an angle (which are triangular in cross-section). Not surprisingly, they found that some shapes work better than others. The field of optimal shapes narrowed further when true-to-life criteria were applied: a serial arrangement for the blades, like that of teeth in the jaw, and a degree of lateral as well as vertical movement, as commonly occurs in chewing.

In the case of the single blades, the most efficient is a symmetrical, notched blade (Fig. 1a), strikingly similar to 'carnassial' teeth that have evolved in several mammal lineages (Fig. 1c). The optimal double-bladed models have three points and two high crests (Fig. 1b), forming a 'protoconoid' that closely resembles the trigonid of simple mammalian lower molars (Fig. 1d, e). This notched triangle is a familiar shape to any student of mammalian evolution, because it evolved early and possibly more than once in mammalian history and is present in many living groups, such as opossums and bats.

Evans and Sanson's study did not address the significant role of crushing in chewing. Their modelling therefore did not produce a 'tribosphenic' tooth shape, characterized on the lower molars by a low basin behind the high trigonid (Fig. 1d, e) into which the largest cusp of the upper tooth fits. Tribosphenic molars perform both slicing and crushing functions, and were present in the ancestors of all living mammals. Although Evans and Sanson focused on cutting alone, the superior efficiency of their protoconoid models, and the evident supremacy of function in determining tooth form, may support the arguments of those who believe that tribospheny evolved two or even three times within early mammals. ■

Anne Weil is in the Department of Biological Anthropology and Anatomy, Duke University, Durham, North Carolina 27708-0383, USA. e-mail: annew@duke.edu

Surface science

View from the edge

David G. Castner

Bombarding a material with an ion beam provides valuable information about its surface composition. Ion guns that fire polyatomic ions can reveal more detailed information and cause less damage to the sample.

Reactions and interactions occur on surfaces, and the composition and structure of a surface are usually quite different from those of the bulk material. So surface properties determine how the material will perform in a given application. Paint adhesion to cars, conversion of petroleum crude oil into gasoline and the wetting of paper by ink are all processes that are driven by surface properties. When a reactant passes over a catalyst, it is the type and arrangement of the catalyst's surface atoms that determine the rate and selectivity of the reactant's conversion to product. Similarly, when a biomaterial is exposed to blood, its surface species determine the composition of the adsorbed blood-protein film and the extent of protein denaturation within the film.

Specialized instrumentation and analysis techniques are needed for characterizing surfaces, as the number of bulk atoms in a material is typically orders of magnitude higher than the number of surface atoms. Thus the bulk signal can easily overwhelm the signal from the surface atoms. The past three decades have seen tremendous

advances in the methods used for surface characterization¹. But no single technique can perform a complete analysis of all types of material — instead, several complementary techniques are used together. Writing in *Applied Surface Science*, Wong *et al.*² and Davies *et al.*³ describe new tools — gold and fullerene cluster-ion sources — to add to this arsenal.

Secondary-ion mass spectrometry (SIMS)⁴ is a leading technique for analysing surfaces. A primary ion beam directed at the material is used to 'sputter', or eject, positively and negatively charged secondary ions from the surface (Fig. 1). In the static mode (in which less than 1% of the surface is struck by incoming primary ions), the secondary ions originate from the outermost nanometre of the sample, providing the desired surface sensitivity. Depending on the type and energy of the incident primary ions and the nature of the material, a wide range of atomic and molecular secondary ions can be produced, which can then be analysed using mass spectrometry. Unfortunately, only a small fraction of the ejected particles are ions, and many of these are atomic species

that give little useful information about the molecular structure of the surface.

Several approaches have been taken to expand the capability of SIMS for surface imaging. For example, the use of 'time-of-flight' mass analysers, with high transmission of ions and parallel mass-detection capability, has improved the efficiency of ion detection⁵, and data analysis has been optimized to extract the maximum information from the complex SIMS spectra and images⁶.

Another focus of effort has been the development of new primary ion sources to produce a higher yield of secondary ions in SIMS experiments. Initially, gas ion sources (such as helium, argon and xenon) were used, as they are relatively easy to build and operate⁴. Increasing the mass of the primary ion (for example, switching from argon to xenon) increased the yield of higher-molecular-weight secondary ions. But the higher mass of the atomic primary ions also increased the amount of damage done to the surface. Another drawback was that atomic gas ion sources are of limited usefulness for imaging experiments because they cannot be focused down to small spot-sizes. Liquid-metal-ion guns⁴ (LMIGs) instead became popular for imaging SIMS applications. Gallium is the element most widely used in LMIGs, but indium and gold are also now used.

The most significant advance came with the development of polyatomic or 'cluster-ion' beams^{7–12}, which achieve an increased yield of high-molecular-weight fragments while keeping sample damage low. The first cluster-ion source (using SF₅) was developed more than ten years ago⁷. Although it showed promise for SIMS analysis, it lacked the intensity needed for widespread use. A successful commercial version⁸ (using SF₅ ions) was developed for spectroscopic analysis, but it cannot be used for high-spatial-resolution imaging because, again, the SF₅ ion beam cannot be focused down to a small spot-size. Also, the lifetime of the SF₅ source is rather short.

Despite the limitations of the first cluster-ion beam sources, they have since developed to become a firm favourite with the SIMS community. Now Wong *et al.*² and Davies *et al.*³ report on two cluster-ion sources that provide high intensity and micrometre-scale spatial resolution, opening up new possibilities for SIMS analysis of materials; the first is a C₆₀ (fullerene) ion source, the second a gold-cluster LMIG. Both produce a significant increase in the yield of higher-molecular-weight fragments from most organic and biological samples. And, in contrast to results obtained by increasing the mass of an atomic ion source, this yield increase does not come with increased damage to the sample surface. In fact, for some samples the damage rate from cluster-ion sources is actually lower.

The C₆₀ source can produce both C₆₀⁺ and

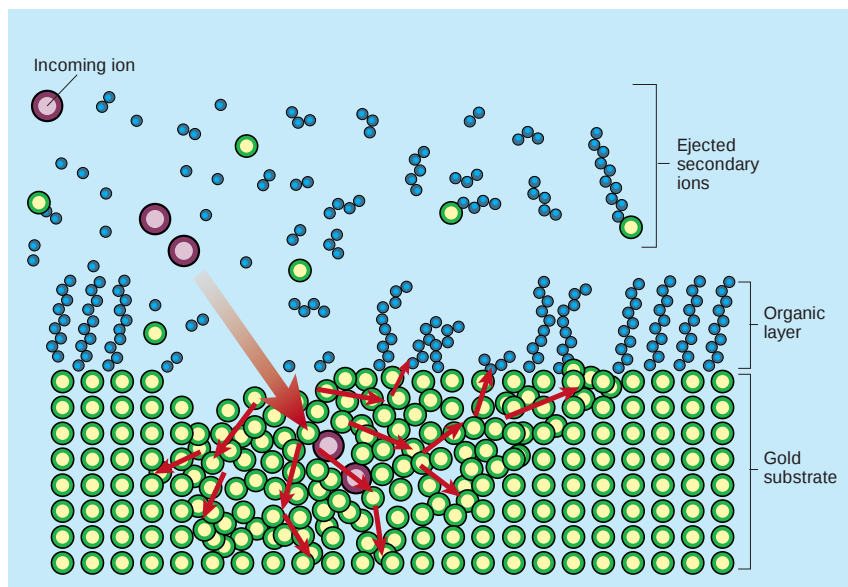


Figure 1 Surface analysis using secondary-ion mass spectrometry (SIMS). In this example, ions are fired into an organic self-assembled monolayer on a gold substrate. The energy deposited in the surface region from the incoming primary ions produces a collision cascade. This results in the ejection of a wide range of atomic and molecular fragments, of which about 1% are ions. Mass analysis of the ejected secondary ions is the key to exploring the structure of the surface. Wong *et al.*² and Davies *et al.*³ have developed primary ion sources that produce a higher yield of secondary ions for analysis.

C_{60}^{2+} ions, focused to a spot size of a few micrometres. The yields of secondary ions obtained using the C_{60} source are typically 50 to 200 times higher than from using gallium LMIGs. The gold-cluster LMIG can produce Au , Au_2^+ and Au_3^+ ions. Secondary-ion yields produced using Au_3^+ ions are typically 10 to 100 times higher than for Ga^+ ions. As the ions from this source can be focused down to smaller spot sizes than with the C_{60} source, it will be the preferred ion source for imaging experiments. On the other hand, the C_{60} source produces a higher yield of molecular fragments from organic and biological samples than the gold source, so it will be preferred for spectroscopic experiments.

Using gallium and indium LMIGs, the spatial resolution for imaging with low-ion-yield molecular fragments is typically limited to a few micrometres: decreasing the pixel size below one micrometre typically causes the number of secondary ions per pixel to drop to essentially zero. But with the increased yield of the gold-cluster LMIG, sub-micrometre spatial resolution could be attainable for molecular SIMS imaging. This will significantly improve the capability of SIMS for imaging the spatial distribution of drugs, proteins and oligonucleotides in a wide range of biomedical devices and sensors.

Although these new cluster-ion sources offer exciting possibilities for SIMS analysis (such as molecular-depth profiling capabilities for organic samples¹¹, something most SIMS analysts had assumed was impossible), several areas still require additional study. At present it is not understood how cluster-ion beam sputtering can simultaneously achieve higher secondary-ion yields and lower damage rates. Several explanations have been proposed: for instance, perhaps it is because the total ion energy is distributed over all the atoms in the cluster, or because there is overlap in the energy cascades produced when the polyatomic ion strikes the surface. Other

questions, such as whether 'increased yield with decreased damage' holds for all types of sample, and whether the concentration of surface species could be quantified, are also unanswered. Further experimental and theoretical research is needed to address these and other issues.

David G. Castner is at the National ESCA and Surface Analysis Center for Biomedical Problems in the Departments of Bioengineering and Chemical Engineering, University of Washington, Seattle, Washington 98195-1750, USA.
e-mail: castner@nb.engr.washington.edu

1. Vickerman, J. C. (ed.) *Surface Analysis — Techniques and Applications* (Wiley, Chichester, 1997).
2. Wong, S. C. C. et al. *Appl. Surf. Sci.* **203–204**, 219–222 (2003).

3. Davies, N. et al. *Appl. Surf. Sci.* **203–204**, 223–227 (2003).
4. Vickerman, J. C. & Briggs, D. (eds) *ToF-SIMS — Surface Analysis by Mass Spectrometry* (SurfaceSpectra/IM Publs, Manchester/Chichester, 2001).
5. Niehuis, E., Heller, T., Feld, F. & Benninghoven, A. *J. Vac. Technol. A* **5**, 1243–1246 (1987).
6. Tyler, B. J. in *ToF-SIMS — Surface Analysis by Mass Spectrometry* (eds Vickerman, J. C. & Briggs, D.) 475–493 (SurfaceSpectra/IM Publs, Manchester/Chichester, 2001).
7. Appelhans, A. D. & Delmore, J. E. *Anal. Chem.* **61**, 1087–1093 (1989).
8. Kotter, F. & Benninghoven, A. *Appl. Surf. Sci.* **133**, 47–57 (1998).
9. Le Beyec, Y. *Int. J. Mass Spectrom.* **174**, 101–117 (1998).
10. Van Stipdonk, M. J., Harris, R. D. & Schweikert, E. A. *Rapid Commun. Mass Spectrom.* **10**, 1987–1991 (1996).
11. Fuoco, E. R., Gillen, G., Wijesundara, M. B. J., Wallace, W. E. & Hanley, L. J. *Phys. Chem.* **105**, 3950–3956 (2001).
12. Walker, A. V. & Winograd, N. *Appl. Surf. Sci.* **203–204**, 198–200 (2003).

Medicine

Smoke signals for lung disease

Anita B. Roberts

A group of proteins that might confer susceptibility to emphysema has been identified. One of them is transforming growth factor- β , and the discovery highlights the many ways of activating this protein in health and disease.

Emphysema is a lung disease that is predicted¹ to become one of the top five causes of death and disability worldwide by 2020. Cigarette smoking is the greatest risk factor for this disease. Despite this correlation, however, only about 15–20% of cigarette smokers develop emphysema. The fact that these susceptible individuals are generally clustered into families hints that there may be certain genes that predispose people to smoking-induced emphysema. On page 169 of this issue, Morris and co-workers² describe how they used clues from gene-array technology and manipulation of the mouse genome to implicate an entirely new group of possible susceptibility genes.

Unlike asthma, in which the flow of air through the lungs is temporarily obstructed,

emphysema is characterized by a progressive airflow restriction that results from permanent enlargement of the lungs' peripheral air spaces and loss of lung elasticity. In smoking-related emphysema these changes are often attributed to the destruction of lung connective tissue through enzymatic degradation of elastin, the main component of the elastic fibres of lung tissue¹.

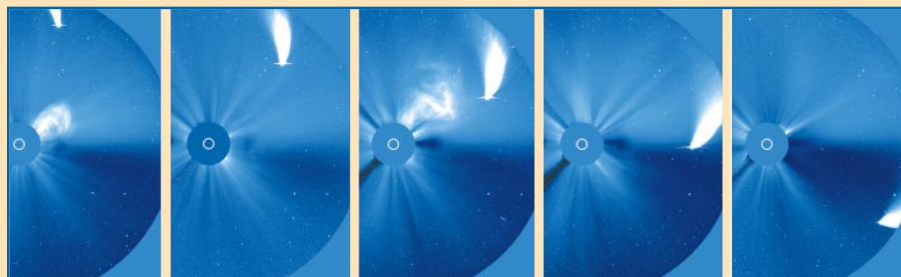
Matrix metalloproteinases (MMPs) are zinc-dependent enzymes that can degrade connective tissues, including elastic fibres^{1,3}. The gene for the MMP12 protein in particular has been considered a strong candidate for conferring susceptibility to emphysema. MMP12 is produced by macrophages — inflammatory cells that infiltrate smokers' lungs — and destroys not only elastin itself,

Solar System

Close encounter of the cometary kind

Last month, the Solar and Heliospheric Observatory, run jointly by the European Space Agency and NASA, enjoyed a spectacular view, as the comet C/2002 V1 (NEAT) approached the Sun at a distance of around 15 million kilometres — roughly one-tenth of the distance from the Sun to the Earth. The comet is seen here over a 69-hour period.

Hourly images were fed back by the large-angle and spectrometric coronagraph, an instrument studying the Sun's corona, or gas halo. To



detect the corona, the coronagraph blots out the disk of the Sun (the star's position is indicated by white circles).

As the comet orbits the Sun, its bright ion tail is deflected by the magnetic field of the solar wind. The

striking feature in the centre picture is a huge eruption of gas from the corona.

Alison Wright

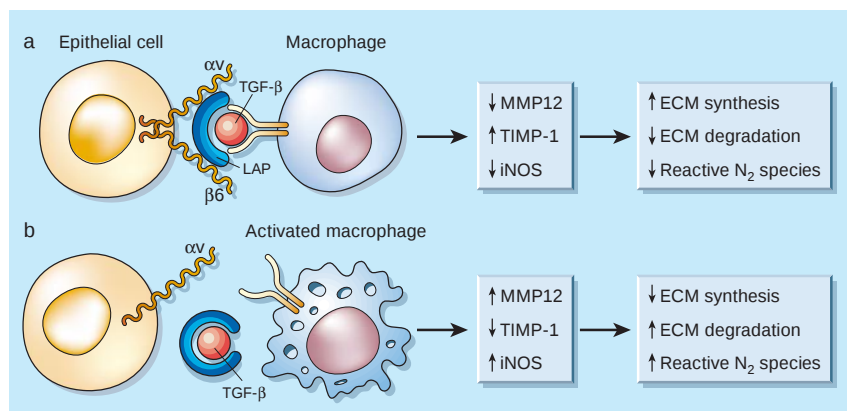


Figure 1 Effects of activating and inhibiting transforming growth factor- β (TGF- β) in the lungs. a, The $\alpha v \beta 6$ integrin protein is found on epithelial cells, such as those that line the lungs. When the latency-associated peptide (LAP) binds to $\alpha v \beta 6$ integrin, it enables TGF- β to bind to its receptor on macrophages⁵ (and fibroblasts; not shown). This results in reduced synthesis by macrophages of the matrix-degrading enzyme MMP12 and of inducible nitric oxide synthase (iNOS). It also results in enhanced production by fibroblasts of components of the extracellular matrix (ECM) and of proteins, such as TIMP-1, that inhibit matrix-degrading enzymes. b, In mice lacking $\beta 6$ integrin, the prevention of TGF- β binding to its receptor results in reduced production of ECM and enzyme inhibitors and increased production of MMPs and iNOS. This protects against pulmonary fibrosis⁵, but, as Morris *et al.* show², the resulting matrix-degrading environment leads to destruction of lung tissue and emphysema.

but also an inhibitor of another type of elastin-degrading enzyme. Moreover, levels of MMP12 are higher than normal in the lungs of mice exposed to cigarette smoke, and deletion of the MMP12 gene protects mice from developing emphysema caused by prolonged exposure to cigarette smoke³. In addition, variations between people in the sequence of their MMP1 and MMP12 genes have been linked to a predisposition to smoking-related emphysema⁴.

Morris *et al.*² have now used genetic manipulation in mice to reveal a new mechanism that leads to abnormal production of MMP12 and to late-onset emphysema that closely mimics the human disease. These authors had previously studied integrin $\alpha v \beta 6$, a protein found on the cell surface. It consists of two subunits, αv and $\beta 6$, and its job is to sense components of the extracellular matrix around cells. The authors found that deleting the $\beta 6$ subunit, which occurs only in epithelial cells (such as those that line the lungs), protects against pulmonary fibrosis (scarring)⁵.

How does this work? It is possibly no accident that the so-called latency-associated peptides of the transforming growth factor (TGF) $\beta 1$ and $\beta 3$ proteins carry a motif that can recognize $\alpha v \beta 6$ integrin, just as matrix components do. When these peptides bind to $\alpha v \beta 6$ integrin in the lungs, they activate extracellular latent TGF- β , which can then bind to receptors on nearby cells, such as macrophages or fibroblasts, and activate signalling^{5,6} (Fig. 1a). This increases the production of matrix proteins by fibroblasts and reduces the production of matrix-degrading metalloproteinases by macro-

phages, leading to abnormal accumulation of matrix components and scarring. So, when $\beta 6$ integrin is deleted, this mechanism for activating latent TGF- β no longer works: matrix metalloproteinases are still generated, matrix production is reduced, and scarring is prevented⁵ (Fig. 1b). Although it has long been recognized that TGF- β can enhance the accumulation of extracellular matrix and promote fibrosis in many tissues, these studies were the first to emphasize the importance of its activation from latent forms in disease.

Morris *et al.* now show that ageing mice lacking $\beta 6$ integrin also develop emphysema. Their disease is characterized by enlargement of the air spaces, infiltration of the lungs by activated macrophages and — as shown by array technology — a 100-fold increase in the amount of MMP12 messenger RNA in the lungs (a measure of how strongly the MMP12 gene is expressed there). MMP12 is clearly necessary for disease onset in this case, because mice lacking both this protein and $\beta 6$ integrin do not develop emphysema.

As TGF- β strongly suppresses MMP12 expression in macrophages⁷, it again seemed likely that mice lacking $\beta 6$ integrin express MMP12 — and develop emphysema — because latent TGF- β in the lungs is not activated (Fig. 1b). This idea is supported by Morris and colleagues' finding that overproduction of either normal $\beta 6$ integrin, or mutant forms of it that can still activate latent TGF- β , decreases MMP12 expression in the lungs of ageing $\beta 6$ -deficient mice, and blocks the development of emphysema. The use of a mutant $\beta 6$ integrin that cannot bind

or activate latent TGF- β does not have these effects. To show more directly that a failure to activate latent TGF- β underlies emphysema in this model, the authors also generated $\beta 6$ -deficient mice that produce permanently active TGF- β . The early signs of emphysema did not develop.

These exciting results broaden the spectrum of factors involved in emphysema to those that activate latent TGF- β in the lung and, by inference, to TGF- β -triggered signalling molecules such as its receptors and intracellular proteins known as Smads⁸. Smad3, for instance, is known to be required for TGF- β to inhibit macrophage activation, and to suppress the production of MMP12 (refs 7, 9). TGF- β also reduces the expression of an enzyme called nitric oxide synthase^{7,9}, presumably leading to decreased levels of reactive nitrogen species — molecules that can amplify inflammation and so contribute to diseases such as emphysema. By contrast, TGF- β activates production of TIMP-1, an enzyme that inhibits matrix metalloproteinases. So, a combination of factors may contribute to the emphysema seen in $\beta 6$ -integrin-deficient mice: an increase in the amount of MMP12 and nitric oxide synthase, and a decrease in TIMP-1 levels.

TGF- β is also involved in suppressing tumour development, and this invites comparisons between the effects of reduced TGF- β signalling in emphysema and lung cancer⁸. Last year, a defect in latent TGF- β -binding protein-4 — a structural component of the extracellular matrix that also affects the targeting of latent TGF- β to the matrix — was shown to result in colon cancer¹⁰. This was again associated with a drastic reduction in TGF- β levels and signalling. Together, these studies draw attention to the often overlooked role of the many tissue-specific, and possibly disease-specific, mechanisms for activating latent TGF- β , which are important not only in disease but also in maintaining normal tissue integrity. Perhaps more importantly, the work of Morris *et al.* provides new potential targets for investigations of candidate susceptibility genes and for the development of drugs to treat or prevent emphysema. ■

Anita B. Roberts is in the Laboratory of Cell Regulation and Carcinogenesis, National Cancer Institute, Building 41, 41 Library Drive, MSC 5055, Bethesda, Maryland 20892-5055, USA.
e-mail: roberts@dce41.nci.nih.gov

1. Barnes, P. J. *Annu. Rev. Med.* **54**, 113–129 (2003).
2. Morris, D. G. *et al.* *Nature* **422**, 169–173 (2003).
3. Hautamaki, R. D. *et al.* *Science* **277**, 2002–2004 (1997).
4. Joos, L. *et al.* *Hum. Mol. Genet.* **11**, 569–576 (2002).
5. Munger, J. S. *et al.* *Cell* **96**, 319–328 (1999).
6. Annes, J. P., Munger, J. S. & Rifkin, D. B. *J. Cell Sci.* **116**, 217–224 (2003).
7. Feinberg, M. W. *et al.* *J. Biol. Chem.* **275**, 25766–25773 (2000).
8. Wakefield, L. M. & Roberts, A. B. *Curr. Opin. Genet. Dev.* **12**, 22–29 (2002).
9. Werner, F. *et al.* *J. Biol. Chem.* **275**, 36653–36658 (2000).
10. Sterner-Kock, A. *et al.* *Genes Dev.* **16**, 2264–2273 (2002).

Optics

Fibres tune in with fluid plug

Appl. Phys. Lett. **82**, 1338–1340 (2003)

A tunable filter has been introduced into an optical fibre by filling hollow channels in the fibre with moveable plugs of liquid. The filter acts as a barrier to infrared radiation within a narrow wavelength band as it passes down the fibre.

Filtering is vital for controlling and routing signals in optical telecommunications. Filters typically consist of an optical grating — a periodic structure that scatters and blocks light at a wavelength determined by the repeat spacing. Such structures have been introduced into optical fibres previously, but the advantage of the filter reported by C. Kerbage and B. J. Eggleton is that it is tunable.

Their grating is made from plugs of organic fluid inserted into six hollow channels surrounding the fibre's core, in which the light travels. These plugs are lined up in a constriction in the fibre, where the light within the core spreads into the channels and interacts with the grating. By warming and expanding the air on either side of the series of fluid plugs, the grating can be compressed so that its periodicity, and thus the wavelength of blocked light, decreases.

Philip Ball

Immunology

Lone rangers

Proc. Natl Acad. Sci. USA **100**, 2604–2609 (2003)

The immune system's T cells circulate between the blood and secondary lymphoid organs such as lymph nodes, continuously surveying the body for foreign molecules. Mark J. Miller *et al.* now reveal the first *in vivo* images of T-cell dynamics within lymph nodes. They find that, contrary to expectation, individual T cells 'go it alone', migrating along independent and seemingly random paths.

Accumulating evidence has suggested that the migration of T cells is directed by attractant molecules called chemokines; from this it was supposed that T cells would stick together, moving in groups towards

chemokine signals. Miller *et al.* have used two-photon imaging — a technique until now used largely by neuroscientists and developmental biologists — to ask exactly how T cells move in live mice.

They find that the cells migrate rapidly, but can also change gear, cycling between periods of low speed when the cells momentarily pause, and high speed (more than 25 μm per minute). The real surprise, however, is that T cells travel alone along random routes, as in the images below. This calls into question the role of chemokines in guiding these cells, and reopens the debate over what, if anything, directs their movement.

Alison Schuldt

Neuroscience

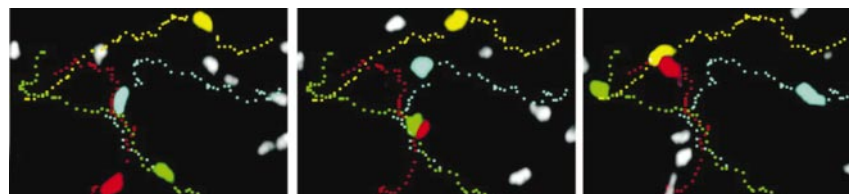
Going to town

Neuron **37**, 877–888 (2003)

People use distinct brain regions to find their way along new routes and along familiar routes. This is the conclusion drawn by Tom Hartley and colleagues, who used magnetic resonance imaging to study brain activity as male subjects navigated their way around two virtual, computer-generated towns. In each town they had to find their way between landmark locations. In one town these were chosen in fixed sequence to form a well-rehearsed route that never changed. In the other town the sequence of landmarks defined new routes that the subjects had not followed before.

Rat studies have shown that the hippocampus is involved in using new routes: animals are thought to form 'cognitive maps' that help them use landmarks to find their way from A to B. But when the same journey is taken again and again, a different brain region, the caudate nucleus, kicks in. Hartley *et al.* found a similar pattern in their subjects: more accurate navigators showed activation of the hippocampus when taking the less-familiar path, whereas the 'head' of the caudate nucleus was active when they followed the well-worn routes. This suggests that accurate navigation depends on selecting the best mental representation for the task in hand. This choice also affects the pattern of brain activity, and may help to explain some of the differences seen between men and women in previous studies of navigation.

Helen R. Pilcher



Far from the crowd: four T cells (false-coloured in different shades) follow independent paths (dotted lines) over periods of minutes.

Biomimetics

Gels got rhythm

Macromolecules doi:10.1021/ma0259618 (2003)

Ryo Yoshida and colleagues report slow, mechanical pulsing of a polymer gel. This unusual behaviour is caused by the Belousov–Zhabotinsky reaction, a chemical process that is analogous to the Krebs cycle for producing energy in biological cells. The gel — poly(*N*-isopropylacrylamide) — is covalently bound to a metal catalyst and immersed in a solution of malonic acid and an oxidizing agent. Periodic changes in the redox properties of the catalyst regularly switch the temperature at which the gel swells or de-swells as a result of changes in the hydrophilicity of the polymer chains.

The oscillatory period of the reaction was controlled by the concentration of the organic acid, the fastest being 75 seconds and the slowest about 830 seconds. The rate of the Belousov–Zhabotinsky reaction limited the amplitude of the mechanical oscillations. Beating could last for two to three hours, depending on the volume of solution. The authors are developing their process further for micro-scale transporter and actuator applications.

Rosamund Daw

Atmospheric science

Quick, quick, slow

Geophys. Res. Lett. doi:10.1029/2002GL015674 (2003)

In July and August 2000, remote-sensing instruments tracked the consequences of two typhoons, Kai-Tak and Bilis, during their passage over oceans in Southeast Asia. Analyses of the data have produced a refined picture of the relationship between sea-surface temperature and wind speed.

Typhoons are driven by the energy from a warm sea-surface, and in turn drag up water from depth as they pass, leaving surface patches that can be up to 6 °C cooler than the surrounding ocean. As I.-I. Lin and co-workers point out, this situation provides a natural experiment. Taking advantage of this, they find that the speed of surface winds drops dramatically over such patches compared with wind speed over neighbouring ocean. This agrees with a previously proposed mechanism of wind-speed modulation at the sea surface. In Lin and colleagues' data, every 1 °C drop in sea temperature typically corresponds to a decrease of about 1 m s^{-1} in wind speed. The cold patches don't take long to warm up again, however, and — most notably — the authors find that events can be played out on comparatively small scales (100–400 km), and quite fast (within a day).

Tim Lincoln

Human footprints in Pleistocene volcanic ash

These ancient tracks are the oldest known to have been made by fully bipedal humans.

We have analysed three fossilized trackways of human footprints in a zeolite-rich pyroclastic flow dated to 385,000–325,000 years ago (kyr), discovered along the western margin of the Roccamonfina volcanic complex in southern Italy. We believe that these tracks are the oldest human footprints found so far and that they were made by hominids who had a fully bipedal, free-standing gait, using their hands only to steady themselves on the difficult descent.

Known locally as ‘devils’ trails’ (and pointed out to us by Marco De Angelis and Adolfo Panarello), the three human trackways are associated with other mammalian tracks in a Middle Pleistocene pyroclastic flow known as the Brown Leucitic Tuff, from the Roccamonfina volcano in north-western Campania, southern Italy.

The stratigraphic location and age of the

trampled layer are supported by lithological comparison with the well-documented radiometric-dated series of the volcanic complex¹. The development of the Roccamonfina volcano can be divided into three main phases²: the first, eruptive phase (630–385 kyr), which ended with a caldera-forming event; a second in which volcanic–tectonic events formed the Brown Leucitic Tuff (385–325 kyr) as well as later pyroclastic flow units; and a third that saw the formation of trachybasaltic lavas and ended about 50,000 years ago.

The tracks (designated here as A, B and C) occur on the surface of a single pyroclastic flow unit that probably belonged to the second phase, and they descend a steep slope that is inclined by up to 80°. The tracks have the same general direction but different stride patterns.

Trackway A is 13.40 m long and consists of 27 footprints across a difference in elevation of 4.26 m; the slope is crossed by a Z-shaped trackway that contains two sharp turns, presumably made in order to negotiate the descent more easily (Fig. 1 a–c).

Trackway B is 8.60 m long and consists of 19 footprints over a difference in elevation of 2.91 m; the slope is crossed in a single straight line that curves roughly 45° to the right (Fig. 1d). The footprints are less evenly spaced after the fifth footprint, as a steeper section is negotiated, and there is evidence of slipping: an occasional handprint is seen on the slope beside the track, placed with an open palm at the side of the trunk (Fig. 1d). By contrast, trackway C (not shown) follows a straight line of 9.98 m, with 10 regular footprints over a smaller incline of 2.56 m, indicating a constant pace with no slipping.

The footprints are about 20 cm in length and 10 cm in width. Using the global average foot-length/stature ratio of 15%, the footprints indicate that the people who made them were no taller than 1.5 m. In some of the prints, the impressions made by the heel and ball of the foot are clear, and there are even small depressions that can be interpreted as toe impressions. The ball impressions have oblique principal axes that run anteromedially to what is thought to be the base of the big toe. In some cases, the central area of the footprint is raised, indicating that the foot was arched. The trackways are narrow, with an average pace of 60 cm and a stride of 120 cm. Although the footprints do not show all of the known features of contouring human bipedalism, there are enough similarities to support the idea that they are indeed human and fully bipedal³.

Although the dating of these trackways is provisional, to our knowledge they are the first human tracks discovered from the Middle Pleistocene period and are therefore the oldest to be found so far.

Paolo Mietto*, Marco Avanzini†, Giuseppe Rolandi‡

*Dipartimento di Geologie, Paleontologie e Geofisica, Università di Padova, 35137 Padova, Italy
e-mail: paolo.mietto@unipd.it

†Museo Tridentino di Scienze Naturali, 38100 Trento, Italy

‡Dipartimento di Geofisica e Vulcanologia, Università di Napoli – FedericoII, 80138 Napoli, Italy

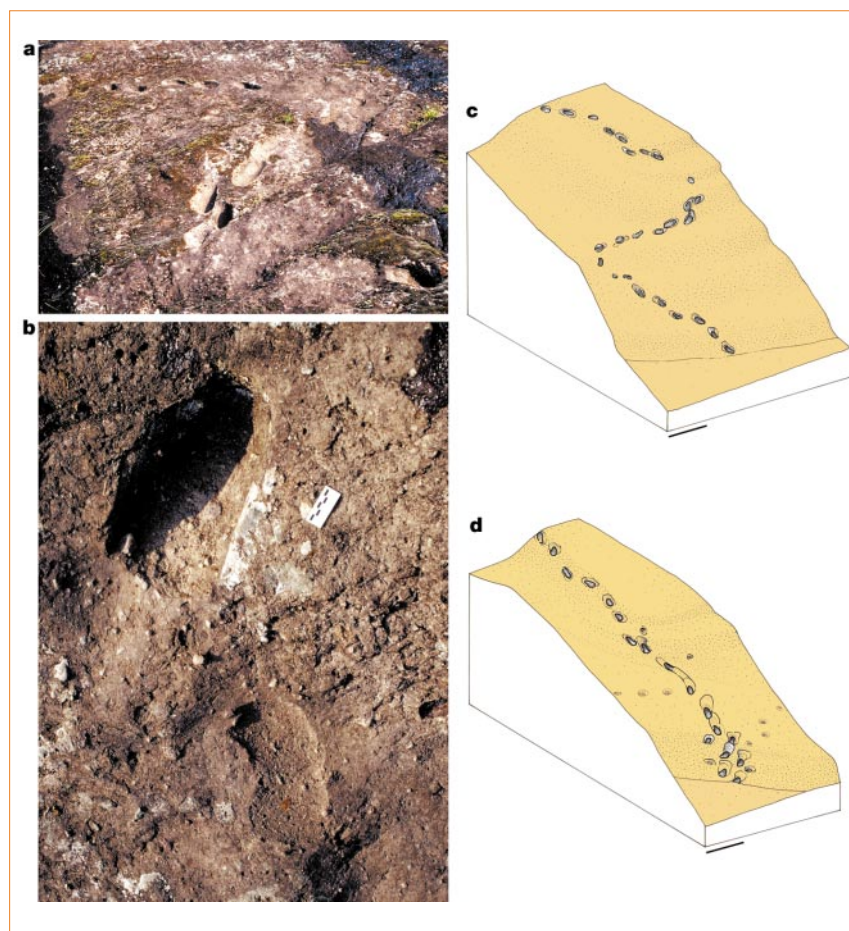


Figure 1 Middle Pleistocene human tracks from the Brown Leucitic Tuff at the Roccamonfina volcano in north-western Campania, southern Italy. **a, b**, Views of the western part of the site showing trail A, which descends an inclined slope (detail enlarged in **b**). **c, d**, Diagrams of trackways A and B, respectively. The trackways indicate the course taken by the trackmaker and therefore his or her direction with respect to the slope. On either side of the final two footprints in trackway B, and also about halfway along it, handprints are evident that probably result from attempts to maintain balance during the difficult descent. Scale bars, 1 m.

1. Ballini, A., Barberi, F., Laurenzi, M. A., Mezzetti, F. & Villa, I. M. *Bollettino Gruppo Nazionale di Vulcanologia* **5**, 533–574 (1989).
2. De Rita, D., Giordano, G. & Milli, S. J. *Volcanol. Geotherm. Res.* **78**, 2267–2288 (1996).
3. Kramer, P. A. & Eck, G. C. J. *Hum. Evol.* **38**, 651–666 (2000).

Competing financial interests: declared none.

Atmospheric science

Ultraviolet light and leaf emission of NO_x

Nitrogen oxides are trace gases that critically affect atmospheric chemistry and aerosol formation¹. Vegetation is usually regarded as a sink for these gases, although nitric oxide and nitrogen dioxide have been detected as natural emissions from plants^{2,3}. Here we use *in situ* measurements to show that solar ultraviolet radiation induces the emission of nitrogen oxide radicals (NO_x) from Scots pine (*Pinus sylvestris*) shoots when ambient concentrations drop below one part per billion. Although this contribution is insignificant on a local scale, our findings suggest that global NO_x emissions from boreal coniferous forests may be comparable to those produced by worldwide industrial and traffic sources.

We measured the amount of NO_x emitted from *P. sylvestris* (NO_x flux) at the SMEAR II station in southern Finland⁴ by enclosing individual young shoots inside a specially equipped chamber^{5,6}. The cover of each chamber was made of ultraviolet-transparent quartz glass (which has a transmittance of over 90% for ultraviolet A and B light) so that the plants were exposed to solar ultraviolet radiation. The chambers were closed 2–3 times every hour for 60 s while the gas concentration inside was being monitored.

We used a mass-balance equation to determine the rate of change in NO_x concentration inside the chamber during a single chamber closure. To calculate the NO_x flux, the solution to the mass-balance equation was fitted to the measured values⁶.

The NO_x concentration increased during closure in an empty chamber, but increased faster when the chamber contained a shoot (Fig. 1a). Ultraviolet radiation also caused a small amount of NO_x emission from the chamber walls, which was taken into account in applying the mass-balance equation.

NO_x flux from the shoots occurred when they were exposed to solar ultraviolet light: when the shoots were removed from their ultraviolet-transparent chambers, the NO_x flux decreased, but was restored when the shoots were replaced (Fig. 1b). We repeated this experiment using two chambers during three sunny days in summer 2001, and found that the measured reduction in NO_x emission after removing the shoots, and the increase after inserting the shoots, were statistically significant on each occasion ($P = 0.00004$).

We confirmed that ultraviolet radiation was inducing NO_x emission from the shoots by replacing the quartz covers on the chambers with ultraviolet-opaque methacrylate

(plexiglass) covers (negligible transmittance at 290–320 nm). NO_x emission decreased from shoots inside chambers with opaque covers, and removing the plants had no effect on NO_x flux (Fig. 1c). Swapping the covers had no effect on photosynthesis, which uses visible light.

The uptake of NO_x by plants increases as its atmospheric concentration increases, so a compensation point⁷ is frequently detected^{2,3,8,9}. NO_x concentrations at the SMEAR II station are usually below 1 part per billion (p.p.b), but in May 2001 they rose to 6 p.p.b.

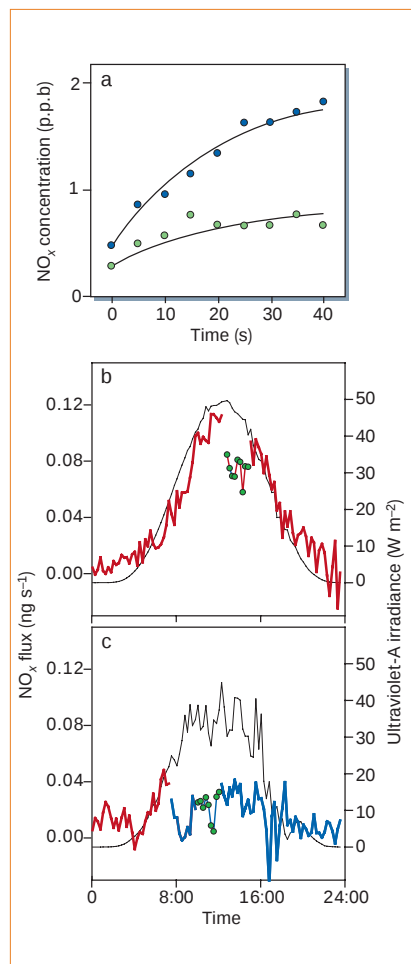


Figure 1 Ultraviolet radiation induces emission of nitrogen oxides from Scots pine (*Pinus sylvestris*) shoots. **a**, Increase in NO_x concentration inside an experimental chamber with an ultraviolet-transparent quartz cover, measured during closure (see text). Blue circles, NO_x concentrations measured inside a chamber containing a single shoot; green circles, NO_x concentrations inside an empty chamber. Curves are solutions to the mass-balance equation used to determine NO_x flux. Shoots were 1 yr old and 10–15 cm high. **b**, Representative experiment showing the change in NO_x flux (red) and ultraviolet-A irradiance (black) in an ultraviolet-transparent plant chamber over 24 h on 28 June 2001. Green circles, NO_x flux in the absence of the plant, which was removed at 12:40 and replaced at 14:50. **c**, Similar experiment, except that the transparent cover was replaced with an ultraviolet-opaque one (blue line) after 07:30 and the shoots were outside the chamber from 10:10 to 12:15.

during an episode when the weather was cloudy and ultraviolet radiation was low (mostly < 10 W m⁻²), causing NO_x deposition. This varied linearly with the ambient concentration, and the compensation point was estimated to be around 1 p.p.b.

As ultraviolet radiation induces NO_x emission from *P. sylvestris* shoots, it must also influence the compensation point. Assuming a linear relationship between exposure to ultraviolet light and NO_x emission, and between the ambient concentration of NO_x and its deposition, the compensation point is above 3 p.p.b. at maximum ultraviolet irradiance. In previous NO_x-exchange studies^{3,8,9}, ultraviolet radiation was excluded either by the chamber material or from the light source, causing the compensation-point estimates to be too low.

We do not yet know the origin of the emitted NO_x. It may come from plant metabolism² or be released from pine-needle surfaces as a result of ultraviolet irradiation, in a reaction similar to the one that occurs at the chamber walls. Gas-phase NO_x is produced from snow by photolysis of nitrate¹⁰ — a reaction that would enable NO_x emissions to recycle reactive nitrogen on foliage.

The NO_x emission induced by ultraviolet radiation was 1 ng m⁻² of pine-needle surface per second under conditions corresponding to a pre-industrial time, when the ambient NO_x concentration was less than 0.5 p.p.b. A flux of this magnitude is sufficient to influence the availability of nitrogen to plants and may affect the atmospheric NO_x budget¹¹.

Pertti Hari*, **Maarit Raivonen***, **Timo Vesala†**, **J. William Munger‡**, **Kim Pilegaard§**, **Markku Kulmala†**

Departments of *Forest Ecology and †Physics, University of Helsinki, 00014 Helsinki, Finland
e-mail: pertti.hari@helsinki.fi

‡Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

§Plant Research Department, Risø National Laboratory, 4000 Roskilde, Denmark

- Seinfeld, J. H. & Pandis, S. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change* (Wiley, New York, 1998).
- Wildt, J., Kley, D., Rockel, A., Rockel, P. & Segsneider, H. J. *J. Geophys. Res.* **102**, D, 5919–5927 (1997).
- Sparks, J. P., Monson, R. K., Sparks, K. L. & Lerdau, M. *Oecologia* **127**, 214–221 (2001).
- Vesala, T. et al. *Trends Heat, Mass Momen. Transfer* **4**, 17–35 (1998).
- Hari, P. et al. *Plant Cell Environ.* **22**, 1297–1301 (1999).
- Altimir, N., Vesala, T., Keronen, P., Kulmala, M. & Hari, P. *Atmos. Environ.* **36**, 19–29 (2002).
- Conrad, R. *Biogeochemistry* **27**, 155–170 (1994).
- Rondón, A. & Granat, L. *Tellus B* **46**, 339–352 (1994).
- Hereid, D. P. & Monson, R. K. *Atmos. Environ.* **35**, 975–983 (2001).
- Honrath, R. E. et al. *J. Geophys. Res.* **105**, 24183–24190 (2000).
- Lerdau, M. T., Munger, W. & Jacob, D. J. *Science* **289**, 2291–2293 (2000).

Competing financial interests: declared none.

Conservation

Reproductive collapse in saiga antelope harems

A common assumption is that breeding in polygynous systems is not limited by the number of males because one male can inseminate many females^{1,2}. But here we show that reproductive collapse in the critically endangered saiga antelope (*Saiga tatarica tatarica*) is likely to have been caused by a catastrophic drop in the number of adult males in this harem-breeding ungulate, probably due to selective poaching for their horns. Fecundity and calf survival are known to be affected by markedly skewed sex ratios³, but in the saiga antelope the sex ratio has become so distorted as to lead to a drastic decline in the number of pregnancies — a finding that has implications both for the conservation of the species and for understanding the reproductive ecology of polygynous ungulates.

S. tatarica tatarica (Fig. 1) is a nomadic inhabitant of the semi-arid rangelands of central Asia and the pre-Caspian. Horns are borne by males and are highly valued in traditional Chinese medicine, which has led

to heavy poaching since the demise of the Soviet Union, when the Chinese–Soviet border was opened. Although there was a marked reduction in the proportion of adult males in the population, fecundity remained high, as was expected for a polygynous ungulate⁴.

Poaching on both sexes increased as the collapse of the rural economy led to local demand for meat⁵, but the sex ratio remained heavily biased towards females. From 1998, saiga populations fell markedly throughout their range, with an average annual rate of decline of 46% (ref. 6). By 2002, the global population was only 50,000 individuals, just 5% of its size 10 years previously⁷, leading to a Critically Endangered listing on the 2002 IUCN Red List and listing on Appendix II of the Convention on Migratory Species.

We collected data on saiga population dynamics in Kalmykia, Russia, over the period 1992–2002. Our data set included the number of *in utero* fetuses in a sample of heavily pregnant females shot during the months of February in the years 1993–2001. During this period, the population size declined rapidly, as did the number of adult males in the population; there was a sudden drop in fecundity in 2001 (Fig. 2a).

Changes in fecundity rates in Kalmykia were not related to population density or climatic variables, unlike in Kazakhstan⁸. The best predictor of female fecundity was the proportion of adult males in the population during the previous year's rut, which explained 96.2% of the variance in fecundity (Fig. 2b). Given that 2001 is a clear outlier and the sample size is small, there is uncertainty attached to this result; however, we could not undertake further analyses as the collection of fecundity data has now ceased for conservation reasons.

The most parsimonious explanation for the drop in fecundity in 2001 is a lack of adult males in the population during the rut. This seems to be a threshold effect, operating at a proportion of adult males in the population between 0.025 (1 adult male per 36 females) and 0.009 (1 adult male per 106 females; Fig. 2b).

The proportion of adult males normally seen in unselectively hunted populations is 0.20–0.25 (ref. 9). The lowest proportions of adult males previously recorded in three other populations were 0.023, 0.077 and 0.053, respectively, in 1992–93, just after sex-biased poaching began⁹, although saiga fecundity remained high in 1993–94 (ref. 4). This supports our inference that the threshold proportion of adult males for reproductive collapse lies below 0.025.

Anecdotal observations of *S. tatarica tatarica* in the 2000 rut point to a complete reversal in reproductive behaviour. Normally, an adult male defends a harem of 12–30 females^{9,10}. In 2000, however, single males



Figure 1 Male saiga antelopes have been so ruthlessly hunted that their breeding and populations have been severely disrupted.

were surrounded by large numbers of females, and dominant females were seen to be aggressively excluding subdominant females from the males: this presumably explains the failure of most first-year females to conceive.

E. J. Milner-Gulland*, **O. M. Bukreeva†**, **T. Coulson‡**, **A. A. Lushchekina§**, **M. V. Kholodova§**, **A. B. Bekenov||**, **I. A. Grachev||**

*Department of Environmental Science and Technology, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
e-mail: e.j.milner-gulland@imperial.ac.uk

†Department of Hunting Management, Elista 358000, Republic of Kalmykia, Russia
‡Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK

§A. N. Severtsov Institute of Ecology and Evolution, Moscow 119071, Russia

||Institute of Zoology, Ministry of Education, Almaty 480032, Kazakhstan

- Bergerud, A. T. in *The Behaviour of Ungulates and its Relation to Management* (eds Geist, V. & Walter, F.) 395–435 (Int. Union Conserv. Nature, Morgues, Switzerland, 1974).
- Haigh, J. R. & Hudson, R. J. *Farming Wapiti and Red Deer* (MosbyYear, St Louis, Missouri, 1993).
- Mysterud, A., Coulson, T. N. & Stenseth, N. C. *J. Anim. Ecol.* **71**, 907–915 (2002).
- Milner-Gulland, E. J., Bekenov, A. B. & Grachev, I. A. *Nature* **377**, 488–489 (1995).
- Robinson, S. *Pastoralism and Land Degradation in Kazakhstan*. Thesis, Univ. Warwick (2000).
- Milner-Gulland, E. J. *et al. Oryx* **35**, 340–345 (2001).
- Sharp, R. *Oryx* **36**, 323 (2002).
- Coulson, T. N., Milner-Gulland, E. J. & Clutton-Brock, T. Proc. R. Soc. Lond. B **267**, 1771–1779 (2000).
- Bekenov, A. B., Grachev, I. A. & Milner-Gulland, E. J. *Mamm. Rev.* **28**, 1–52 (1998).
- Sokolov, V. E. & Zhirmov, L. V. (eds) *The Saiga: Phylogeny, Systematics, Ecology, Conservation and Use* (Russian Acad. Sci., Moscow, 1998; in Russian).

Competing financial interests: declared none.

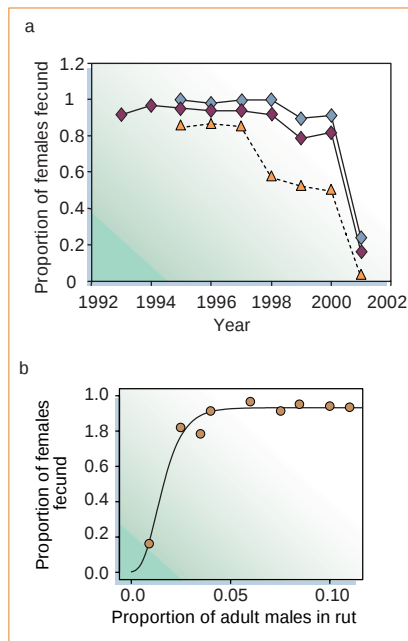


Figure 2 Changes in the fecundity of saiga antelopes (*Saiga tatarica tatarica*) in Kalmykia, Russia. **a**, Proportion of adult (blue diamonds) and juvenile (under 1 year old; triangles) females that were fecund in each year from 1993 to 2001; means are indicated by red diamonds. A mean of 69 females (range, 24–103) was sampled each year. **b**, Relationship between the proportion of adult males in rutting herds in the months of December during the years 1992–2000, and female fecundity the following spring, fitted as a Gompertz regression; the fit explains 96.2% of the variance. The proportion of younger males and adult females also explained over 90% of the variance in fecundity, which is not surprising as these are proportional values. None of the other potential covariates tested gave a satisfactory fit.

Mantle deformation or processing artefact?

Measurements of shear-wave splitting represent an important tool for determining seismic anisotropy and for quantifying deformation processes in the Earth. Wookey *et al.*¹ claim to have observed significant seismic anisotropy in the mid-mantle between the Tonga–Kermadec subduction zone and Australia. We argue that their results are likely to be methodological artefacts and that the available data can be explained by moderate anisotropy in the upper mantle close to the seismograph stations. The lack of evidence for anisotropy in the mid-mantle nullifies any related geodynamic inferences.

Elastic anisotropy results in different propagation velocities of shear (S) waves with different polarizations. Wookey *et al.*¹ report 2–7 seconds of splitting for 15 of 35 S-waves from deep earthquakes, recorded in Australia at epicentral distances of 24–59°. Their observation that SH (motion perpendicular to the wave-propagation plane) generally leads SV (motion within the wave-propagation plane) is interpreted as transverse isotropy (or radial anisotropy with a vertical slow axis) along subhorizontal parts of the wavepaths. This type of anisotropy is well known at the bottom of the lower mantle², but not in the mid-mantle.

To measure S-wave splitting, Wookey *et al.*¹ adopted a method that is normally used to study mantle anisotropy using SKS waveforms, which travel through the Earth's core and become pure SV waves upon re-entering the mantle. The presence of mantle anisotropy beneath a station causes SKS to split into a fast and slow quasi-S-wave, the amount of splitting allowing quantification of the anisotropy. This technique requires S-waves that are steeply incident at the receiver, resulting in almost pure SV motion

on the radial component and negligible interactions with near-receiver structure.

The main obstacle to the approach of Wookey *et al.*¹ is the presence of large-amplitude shear-coupled compressional (P) waves throughout their data set. At epicentral distances of less than 60°, these reverberation phases represent a large percentage of the energy recorded on the radial component^{3,4} (Fig. 1). Wookey *et al.*¹ mistakenly invert this mixture of direct SV and shear-coupled P-waves as pure SV motion, rendering their splitting measurements meaningless.

Through comparison of SH and SV waveforms, which is a standard technique to investigate transversely isotropic media, we test the proposal put forward by Wookey *et al.*¹. To avoid contamination of the SV signal by P phases, we decompose the P and SV wavefields through further rotation of radial and vertical components⁵ (Fig. 1). Examining all of the available data for the stations, we found 128 recordings with qualities that are comparable to those of the 35 used by Wookey *et al.*

Our results are very different from those of Wookey *et al.* Splitting times are consistently small and strongly station-dependent, ranging from almost zero at station CTAO to 1.3 s at WRAB and TAU. At stations NWAQ and CAN, averages amount to 0.4 and 0.7 s, respectively. Wherever splitting is detectable, the fast axis coincides with SH. Owing to their shallow incidence in the upper mantle, S-waves, as opposed to SKS, are sensitive to radial anisotropy. The relatively small splitting times can therefore be explained entirely by a radially anisotropic uppermost mantle⁶ and do not require anisotropy at mid-mantle level.

Joachim Saul*, Lev Vinnik†

*GeoForschungsZentrum Potsdam, Telegrafenberg, 14473 Potsdam, Germany
e-mail: saul@gfz-potsdam.de

†Institute of Earth Physics, Moscow 123995, Russia

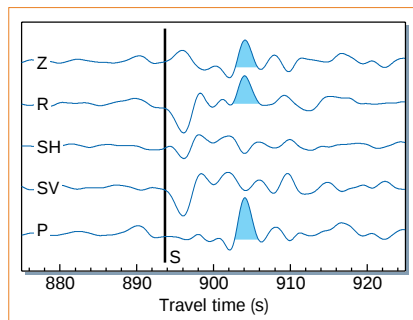


Figure 1 Example of a recording used by Wookey *et al.*¹ (obtained on 13 April 1995 at station NWAQ), who analysed traces labelled R and SH and obtained a splitting time of 4.35 s. A shear-coupled P-wave (shaded) severely contaminates the radial (R) component. SV and SH have very similar waveforms and do not show any difference in travel time. All traces are plotted at the same amplitude scale.

Wookey *et al.* reply — Saul and Vinnik's comment raises an interesting point concerning the influence of shear-coupled P-waves (such as SPmP) on shear-wave-splitting analysis. We used a method^{1,2} that applies a time window to the data, so P-wave energy on the horizontal components outside this window will have no effect on the analysis. As all but one of our events were recorded at stations located on thick (40–50 km) continental crust³, the Moho multiple arrivals to which we refer will

generally be separated (>8 s for SPmP) from the main S-wave arrival. Error can occur when P-wave energy falls within the analysis window.

We have re-analysed our data, applying a wavefield-decomposition technique⁴ before the shear-wave splitting analysis. The range of results changed from 0.6–7.1 s to 0.6–6.2 s, with ten results showing splitting greater than 1.5 s. This is more splitting than can be explained solely by VTI (transverse isotropy with a vertical axis of symmetry) in the upper mantle and still supports our conclusion of mid-mantle anisotropy.

The mistake that Saul and Vinnik have made is to assume a VTI medium and simply to look for splitting between the radial and transverse components. Although our results show a near-VTI symmetry on average ($\varphi_{AVE} = 95.3^\circ$), there is plenty of scatter in the data. The lack of apparent splitting on the components of the seismogram (Fig. 1) is meaningless — the arrivals on the radial and transverse components are combinations of the off-axis fast and slow shear waves.

Our own analysis of this event using the wavefield-decomposed seismogram gives an even more convincing result than previously. We recover a lag time of 3.55 ± 0.2 s and a fast direction of $22 \pm 2^\circ$. Rotated into the correct orientation, the traces show a clear split between similar shear waveforms; if the phase were truly not split (that is, no anisotropy), no such orientation would be possible.

This result shows more splitting than can be explained solely by VTI in the upper mantle. The *a priori* assumption of VTI media means that the methodology of Saul and Vinnik will be unable to detect any splitting with a fast direction that is neither horizontal nor vertical: even a small deviation will cause cross-contamination of the components.

Furthermore, their simple inspection of the waveforms is a qualitative method, relying on a subjective pick of the two phases, with no way of assessing errors. Our method produces a quantitative measurement of the error, allowing us to reduce our initial data set of 290 measurements to the 35 that are constrained sufficiently to draw robust inferences. We are left with a high-quality data set that we consider to be compelling evidence of mid-mantle anisotropy.

James Wookey*, J.-Michael Kendall*, Guilhem Barruol†

*School of Earth Sciences, University of Leeds, Leeds LS2 9JT, UK

e-mail: m.kendall@earth.leeds.ac.uk

†CNRS, Université Montpellier, 34095 Montpellier, Cedex 05, France

1. Wookey, J., Kendall, J.-M. & Barruol, G. *Nature* **415**, 777–780 (2002).
2. Silver, P. G. & Chan, W. W. *Nature* **335**, 34–39 (2002).
3. Clitheroe, G., Gudmundsson, O. & Kennett, B. L. N. *J. Geophys. Res.* **105**, 13697–13713 (1998).
4. Bostock, M. G. *J. Geophys. Res.* **103**, 21183–21200 (1998).

Detrimental effects of sanctions on human altruism

Ernst Fehr* & Bettina Rockenbach†

* University of Zürich, Institute for Empirical Research in Economics, Blümlisalpstrasse 10, CH-8006 Zürich, Switzerland

† University of Erfurt, Nordhäuser Straße 63, D-99089 Erfurt, Germany

The existence of cooperation and social order among genetically unrelated individuals is a fundamental problem in the behavioural sciences. The prevailing approaches in biology and economics view cooperation exclusively as self-interested behaviour—unrelated individuals cooperate only if they face economic rewards or sanctions rendering cooperation a self-interested choice. Whether economic incentives are perceived as just or legitimate does not matter in these theories. Fairness-based altruism is, however, a powerful source of human cooperation. Here we show experimentally that the prevailing self-interest approach has serious shortcomings because it overlooks negative effects of sanctions on human altruism. Sanctions revealing selfish or greedy intentions destroy altruistic cooperation almost completely, whereas sanctions perceived as fair leave altruism intact. These findings challenge proximate and ultimate theories of human cooperation that neglect the distinction between fair and unfair sanctions, and they are probably relevant in all domains in which voluntary compliance matters—in relations between spouses, in the education of children, in business relations and organizations as well as in markets.

Human societies are characterized by an unprecedented division of labour supported by a myriad of social and economic exchanges and compliance with social norms^{1,2}. Norm compliance and the exchange of favours, goods and services pervade every aspect of human life. They shape the interactions in families, neighbourhoods, schools, firms, markets and politics. The crucial feature of any exchange is that the parties involved have to trust each other. If the exchange partners had doubts about the other's reliability most exchanges would not take place. In fact, much of the economic backwardness in the world can plausibly be explained by a lack of mutual confidence inhibiting cooperation in the production and the exchange of goods and services^{3,4}. Yet, what ensures that exchange partners trust each other? In modern societies one solution is to conclude a legally enforceable contract that regulates all aspects of the exchange—in particular, the sanctions imposed on those who breach the contract. If the punishment for breaching the contract is large enough, it is in the self-interest of the involved parties to fulfil their obligations because otherwise they will be heavily punished.

For millennia, however, humans have not been able to rely on contracts because there were no impartial courts enforcing voluntary agreements. Even today it is typically not possible to regulate transactions in every detail, and frequently the courts are unable to verify who violated an agreement. Therefore, even in modern societies the overwhelming majority of all social and economic exchanges are not based on complete and legally enforceable contracts but on implicit agreements and social norms lacking explicitly specified sanctions for non-compliance^{5,6}. In such situations there are ample opportunities for cheating the exchange partner to one's own advantage. It is the ubiquity of such cheating opportunities that renders altruistic cooperation important. Altruistic cooperators are willing to cooperate, that is, to abide by the implicit agreement, although cheating would be economically beneficial for them. Much recent research indicates that altruistic cooperation is an important behavioural force^{7–11}.

However, the prevailing theoretical approaches in biology^{12–17} and economics^{18,19} neglect costly altruistic behaviour that does not yield future economic benefits for the altruist. Individuals are supposed to do others a favour only if they can themselves reap direct^{12,13} or indirect^{14–17} future benefits from such others. Likewise, individuals are supposed to cheat on agreements unless they are

sufficiently rewarded for compliance or sanctioned for non-compliance. This axiom of self-interested behaviour implies that the only remedy for the compliance problem lies in the provision of sufficient rewards or sanctions. Yet, if these incentives for cooperation undermine altruistic behaviour the remedy may do more harm than good because it aggravates the compliance problem^{5,20–23}. Influential social scientists have hypothesized that economic incentives might undermine altruism^{20,21} but so far it has been impossible to provide compelling evidence from uncontrolled field studies.

Capturing trust and altruistic cooperation

We examine the question of how sanctions intended to prevent cheating affect human altruism in a sequentially played social dilemma experiment with real monetary stakes. In each trial of the experiment two mutually anonymous subjects are involved—one subject in the role of investor, the other one in the role of trustee. First, the investor has the chance of choosing a costly trusting action. Then the trustee is informed about the investor's action and he can honour the investor's trust by taking a costly cooperative action. The payoff rules of the experiment ensure that if the investor chooses a trusting action and the trustee responds cooperatively both players increase their monetary payoff. However, the trustee also has the option of not honouring the investor's trust. In this case the trustee saves the costs of cooperation but the investor is worse off than if he had not been trustful. Because a cooperative action is costly for the trustee the trustee is tempted not to honour the investor's trust. Yet, if the investor expects that the trustee will not honour his trust the investor will not choose a costly trusting action in the first place. Thus, the subjects are caught in a dilemma because trust and cooperation is beneficial for both subjects but the trustee faces the temptation not to cooperate and, therefore, the investor is tempted not to trust.

To examine potential detrimental effects of sanctions on human altruism we implement two experimental conditions—an incentive condition and a trust condition. In the trust condition, cooperation of the trustee cannot be achieved by incentives because there are no rewards or sanctions available to the investor. The trustee's cooperation in this condition provides, therefore, a baseline measure of altruism because—in the absence of economic incentives—purely selfish trustees will never cooperate. In the incentive

condition, the investor has the option of fining a trustee who has not cooperated sufficiently. The incentive condition informs us, therefore, whether there are detrimental effects of sanctions on the trustee's altruism. In particular, if the amount of cooperation is higher in the absence of the sanction that would be evidence of a detrimental effect.

The details of the trust condition are as follows. Both the investor and the trustee receive an endowment of ten money units (MUs). The investor has the option to trust his partner by sending between zero and ten MUs of his endowment to the trustee. The experimenter then triples any amount sent so that if the investor sends, say, 10 MUs the trustee receives 30 MUs. The tripling of the investment mimics a situation where the trustee has superior productive opportunities for the use of economic resources. If the investor transfers money to the trustee he also has to specify a 'desired back-transfer' that can be any amount between zero and the tripled transfer. In the previous example, for instance, where the investor sent 10 and the trustee received 30 MUs, the desired back-transfer can be any amount between zero and 30 MUs. Once the investor has made his investment the trustee is informed about the amount sent and the desired back-transfer. Then the trustee has the option to cooperate by choosing the actual level of the back-transfer. The trustee is free to send any amount between zero and the tripled transfer—in our example, up to 30 MUs—back to the investor. The back-transfer is not tripled by the experimenter—if the trustee sends back, say, 8 MUs the investor receives exactly 8 MUs.

The final payoff of the investor is given by his endowment of 10 MUs minus the transfer to the trustee plus the actual back-transfer of the trustee. The payoff of the trustee is given by the endowment of 10 MUs plus the tripled transfer minus the actual back-transfer. Thus, if there is no trust and cooperation the investor's transfer and the trustee's back-transfer are zero. In this case, both earn just their endowment of 10 MUs. This is the predicted outcome if both subjects are fully selfish and the investor anticipates the trustee's selfishness. In the trust condition, a selfish trustee will never pay back anything and, therefore, a selfish investor, who anticipates the trustee's behaviour, will transfer nothing. Yet, trust and cooperation can render both subjects better off. If the investor trusts fully and sends 10 MUs and the trustee's actual back-transfer is 20 MUs the investor earns 20 MUs, and the trustee earns

his endowment of 10 plus the tripled transfer of 30 minus the back-transfer of 20, which in total also gives 20 MUs.

The incentive condition is exactly identical to the trust condition except for one feature: the investor can impose a fine of four MUs on the trustee if less than the desired amount has been sent back. Yet, the investor can also refrain from imposing the fine. The trustee need not be fined even if he pays back less than the desired amount. At the time when the investor chooses the transfer and the desired back-transfer he also has to specify whether to impose a fine for the case that the trustee's actual back-transfer is lower than the desired one. Thus, when the trustee makes his decision in the incentive condition he knows whether the investor has chosen the fining option. To avoid evocative language the experimental instructions do not include value-laden terms like fine or punishment. Instead, the fine is described, more neutrally, as a deduction from the trustee's payoff. The fine affects only the trustee's payoff by reducing his earnings by 4 MUs in the case of an insufficient back-transfer. The fine does not affect the investor's payoff directly. However, the investor can use the fine to affect the trustee's back-transfer. In the incentive condition, a selfish trustee who does not face a fine will never pay back anything. Only if he faces a fine will he be ready to pay back up to 4 MUs. For instance, if the desired back-transfer is 3 MUs the trustee is better off by actually paying back 3 MUs instead of paying the fine of 4 MUs. Therefore, if the world were populated only by selfish subjects we should observe that all investors impose a fine and that only if a fine is imposed will the trustees pay back a positive amount. In such a world the available sanction would be used to enforce the cooperation of selfish trustees, and altruistic cooperation would be absent.

All the interactions in the experiment took place anonymously and each pair of subjects played the experiment only once. This means that repeated interactions^{12,13} or reputation formation^{14–17} cannot account for positive back-transfers. In the absence of a fine, positive back-transfers can, therefore, be characterized as altruistic cooperation. Likewise, if the investor imposes a fine, back-transfers exceeding the fine of 4 MUs represent altruistic cooperation because selfish trustees will never pay back more than 4 MUs.

Sanctions and altruistic cooperation

In our first experiment, 24 pairs of subjects participated in the trust condition and 45 pairs in the incentive condition. To examine the motives behind the investors' choices in more depth we later conducted an additional experiment with 50 pairs in the incentive condition. Figure 1 shows the behaviour of trustees in the two conditions of our first experiment. In contrast to the self-interest hypothesis the trustees paid back substantial amounts of money in all conditions. In addition, back-transfers were increasing in the investors' transfer, indicating the reciprocal nature of human altruism. This kind of altruism should not, however, be confused with the notion of "reciprocal altruism" as defined by Trivers¹². In our experiment there are no repeated interactions so that trustees cannot hope to receive future rewards from their current altruistic acts. 19 of the 24 trustees (79%) in the trust condition paid back more than zero. 29 of the 45 trustees (64%) in the incentive condition paid back more than 4 MUs. This confirms that altruism is an important force in our setting. More importantly, however, there were striking differences in trustees' altruism across conditions (Fig. 1). For any interval of the investors' transfers the trustees' back-transfers were highest when the investor voluntarily refrained from the fine in the incentive condition and lowest when the investor imposed the fine. The back-transfers in the trust condition were always at an intermediate level. When the investor voluntarily refrained from the fine no trustee chose a back-transfer of zero and 47% (7 out of 15 subjects) chose a back-transfer of 15 MUs or more, yielding an average back-transfer of 12.5 MUs (Table 1). In contrast, if a fine was imposed, 33% (10 out of 30 subjects) paid back nothing and only 13% (4 out of 30 subjects) paid

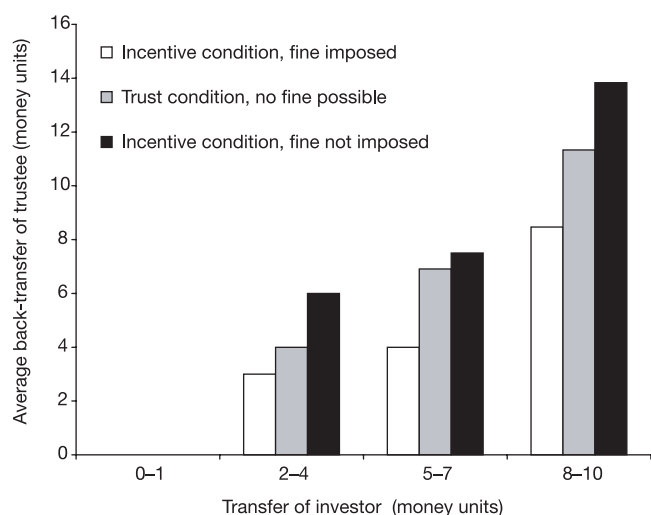


Figure 1 The average back-transfer of the trustees plotted as a function of the investors' transfers. In the trust condition and the incentive condition the back-transfers of the trustees increase with the investors' transfers—irrespective of whether the fine is imposed or not. If the investors impose a fine in the incentive condition, trustees reduce their back-transfers, indicating a detrimental effect of the incentive on altruistic cooperation. Trustees' back-transfers are highest if the investor deliberately refrains from imposing the fine.

Table 1 Average behaviour and payoffs of investors and trustees

	Trust condition	Incentive condition, fine chosen	Incentive condition, no fine chosen
Investment	6.5	6.8	8.7
Desired back-transfer as a percentage of tripled investment	59.9	67.4	63.7
Actual back-transfer	7.8	6.0	12.5
Actual back-transfer as a percentage of tripled investment	40.6	30.3	47.6
Actual back-transfer as a percentage of desired back-transfer	74.4	54.5	74.1
Investor's payoff	11.3	9.2	13.8
Trustee's payoff	21.8	22.4	23.5
Number of observations	24 pairs	30 pairs	15 pairs

back 15 MUs or more, causing an average back-transfer of only 6 MUs. This difference in back-transfers is significant (Mann–Whitney test, $z = -2.988$, $P = 0.003$, two-tailed), providing a first indication that the imposition of the fine had detrimental effects on altruistic behaviour.

One reason for the higher back-transfers in the absence of fines could be that those investors who did not impose a fine transferred higher amounts to the trustees (Table 1). Investors who imposed a fine transferred 6.8 MUs on average whereas those who refrained from fining transferred 8.7 MUs (Mann–Whitney test, $z = -1.961$, $P = 0.050$, two-tailed). Trustees tended to pay back more money when they had received higher transfers (Fig. 1), so the higher back-transfers in the absence of the fine could be due to the higher transfers of the investors. To control for this effect we computed the actual back-transfer as a percentage of the tripled investment (Table 1). We find that if the fine was imposed the trustees paid back 30.3% of the tripled investment, and if investors voluntarily refrained from fining, significantly more, 47.6%, was paid back (Mann–Whitney test, $z = -1.930$, $P = 0.054$, two-tailed). A similar picture emerges if we examine how much the trustees paid back relative to the desired back-transfer. If the fine was imposed the trustees paid back only 54.5% of the desired level whereas in the absence of the fine 74.1% of the desired level was paid back (Table 1).

The negative impact of sanctions on altruistic responses is also supported by a regression analysis, which controls for the investor's transfers and the desired back-transfers. Keeping these other factors constant, using the fine in the incentive treatment reduced the back-transfers significantly by 4.56 MUs ($t = 2.380$, $P = 0.020$, two-tailed). As a consequence, investors who did not use the fine earned significantly more money (Mann–Whitney test, $z = -2.283$, $P = 0.022$): 13.8 MUs in the absence of the fine and only 9.2 MUs when the fine was imposed (Table 1). The imposition of the fine reduced the back-transfers by 4.56 MUs, so it is tempting to assume that trustees who decided to pay back less than the desired amount—and hence had to pay the fine—reduced the back-transfer by the amount of the fine. This was, however, not the case. Those trustees who were actually fined paid back 2.94 MUs on average, whereas those who transferred back the desired amount, so that they did not have to pay the fine, paid back 9.5 MUs on the average. The median back-transfer of those who actually had to pay the fine was zero MUs.

Why do so many investors in the incentive treatment—two-thirds (30 out of 45)—threaten to sanction the trustee if the threat is associated with much lower back-transfers? We hypothesized that there are two potential reasons for this behaviour. First, preferences for strong reciprocity^{24,25}, which have been shown to prevail in many situations^{8,9,11}, could be the proximate cause. A strong reciprocator is willing to sacrifice resources for rewarding behaviour that is perceived as kind or fair and for punishing behaviour perceived as hostile or unfair, even if reciprocation is costly and provides no present or future material benefits whatsoever²⁴. The variance in trustees' back-transfers is considerable, so the investor always faces the risk of being cheated. Therefore, strong reciprocators will impose a fine in case of malfeasance even if it reduces their overall

earnings. Second, investors may simply not have anticipated the detrimental effects of the incentive on the back-transfers.

To discriminate between these two hypotheses we conducted a second experiment with 50 pairs of subjects in the incentive treatment. In this experiment, we informed the investors about the behaviour of the trustees in the first incentive treatment. We gave them two scatter diagrams that showed the trustees' back-transfers for every level of the investors' transfers. One diagram depicts the trustees' behaviour when they faced the fine, the other one shows trustees' behaviour in the absence of the fine. If the investors' preferences for fining were caused by false expectations about the effects of the fine they should change their decisions when informed about the true effects of the fine. This did, however, not occur. Again, roughly two-thirds of the investors (34 out of 50) preferred to fine the trustees in case of non-compliance. This suggests that the desire to punish cheaters is the reason for the frequent use of the fine.

Why do sanctions reduce altruistic cooperation?

All the other qualitative findings observed in the first incentive treatment were also replicated in our second experiment. In particular, we again observed a sizeable increase in the back-transfers if the investors voluntarily waived the possibility of imposing the fine. How can we explain this behaviour by trustees? The notion of strong reciprocity can provide an answer to this question, too. First, refraining from the threat of fining, although the threat is available, could itself be perceived as a fair act, which

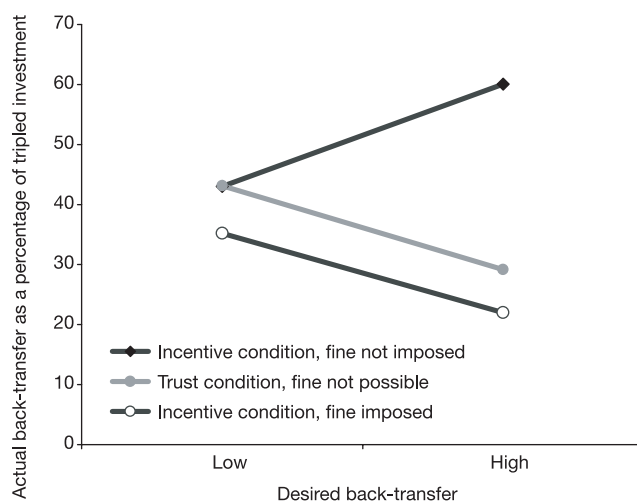


Figure 2 Actual back-transfer as a percentage of tripled investment plotted as a function of the investors' desired back-transfers. The desired back-transfer is categorized as low if, in the case that the trustee pays this back-transfer, the investor would earn the same or less than the trustee. It is categorized as high if the investor would earn more than the trustee. At low desired back-transfers the sanctioning threat reduces actual back-transfers in the incentive treatment but the effect is not statistically significant. At high desired back-transfers the sanctioning threat has a large negative impact on the actual back-transfers in the incentive condition, indicating that if sanctions are used to achieve a distributional advantage they strongly undermine altruistic cooperation.

induces the trustees to increase their cooperation. Second, attempts to use the sanction to enforce an unfair distribution of income may be perceived as hostile acts, inducing the trustees to reduce cooperation.

A back-transfer of two-thirds of the tripled transfer equalizes the trustee's and the investor's payoff. Therefore, we classified desired back-transfers of two-thirds or less of the tripled transfer as low, and desired back-transfers above two-thirds of the tripled transfer were classified as high. We find that for low desired back-transfers the imposition of the fine decreases the actual back-transfers in the incentive condition from 43% to 35.2% of the tripled transfer (Fig. 2). This is evidence that the fine itself diminishes altruistic responses but the effect is not significant (Mann–Whitney test, $z = -0.520$, $P = 0.603$, two-tailed). A significantly large effect occurs, however, if the fine is combined with high desired back-transfers. In this case, the imposition of the fine leads to back-transfers of only 22% of tripled investments whereas voluntarily refraining from fining elicits back-transfers of 60.1% (Fig. 2, Mann–Whitney test, $z = -2.262$, $P = 0.024$, two-tailed). In fact, if the fine is imposed and the desired back-transfers are high, 46% of the trustees give back nothing and the average back-transfers are only 3.82 MUs, which suggests that altruistic responses have largely vanished.

Previous research on 'public good' experiments has shown that altruistic punishment is a highly effective means of enforcing cooperation²⁶. Why do the sanctions in the public good context enhance cooperation but reduce cooperation in our context? Our results suggest that the moral legitimacy of the sanction is a crucial factor. In the public good context the punishment of 'free-riders' is an altruistic act that is considered as morally legitimate. Therefore, the disciplining effect of the sanctions is not undermined by a decline in altruistic cooperation. In the present context, punishment serves the punisher's self-interest and, if it is used to enforce an unfair payoff distribution in favour of the punisher, it decreases altruistic responses. This interpretation is also supported by the following fact. If the fine is imposed in the incentive condition the actual back-transfers as a percentage of tripled investments are significantly lower when the desired back-transfers are high than when the desired back-transfers are low (Mann–Whitney test, $z = -1.85$, $P = 0.064$, two-tailed).

In human societies, social order and cooperation rely on both the use of rewards and sanctions, which ensures the compliance of self-interested actors, and on the presence of people willing to perform altruistic acts. However, as our experiments indicate, sanctions intended to deter non-cooperation may backfire because they undermine altruistic cooperation. Whereas altruistically motivated sanctions for the benefit of the group enhance cooperative behaviour, sanctions that are imposed to enforce an unfair distribution of resources have the opposite effect. This is in contrast to prevailing approaches in economics, biology and behavioural ecology, which predict cooperation-enhancing effects of sanctions, regardless of the moral legitimacy or purpose of each sanction.

Our results, however, do not imply that economic incentives generally have negative motivational effects. There is evidence indicating that rewarding incentives and moralistic sanctions in repeated interactions have large positive effects on cooperation²⁷. This suggests that economic incentives cause mainly negative effects on altruistic cooperation if they come in the form of sanctions and if they are associated with greedy or selfish intentions. We believe that these insights are important for everyone who needs the voluntary compliance of other people. □

Methods

The experiments were conducted with 238 students at the University of Bonn. They took place in a large lunch room for students where we recruited subjects on the spot to

participate in the experiment. At noon the lunch room is visited by hundreds of students from various disciplines. Each pair of subjects played the game just once and no subject could participate in the experiment for a second time. At registration, subjects received an instruction sheet (available from the authors upon request) that explained the rules of the game, the payoff structure, the random and anonymous matching with another student subject recruited at the same occasion, the conversion rate of experimental MUs into German Marks and the payment mode. Subjects were privately paid immediately after the experimental session. After having read the instructions, each subject was randomly assigned to one of two roles (investor or trustee) and anonymously matched with a subject in the other role. Investors received a blank decision sheet in which they had to fill in the transfer and the desired payback amounts. In the incentive condition, they additionally had to mark one of two options 'no deduction' and 'a deduction of 4 MUs' for the case where the trustee would pay back less than the desired amount. While making their decisions subjects were seated in a polling booth that we had brought to the lunch room. After an investor had made his three decisions a monitor checked whether all the necessary information was on the decision sheet. Then the monitor handed the decision sheet to the trustee with whom the investor was matched. The trustees had just one decision to make, namely, to specify the back-transfer to the investor. After completion the decision sheet was handed to the monitor who controlled for completeness and calculated the payoffs. Immediately after their game was over subjects were paid. Special care was taken to ensure the anonymity of the subjects and the privacy of the payment procedure. Each subject received a flat fee of DM3 (€1.53) for participation and each MU earned was rewarded with DM0.5 (€0.26). On average an investor earned DM8.45 (€4.32) and a trustee earned DM14.23 (€7.28). On average half an hour elapsed between registration and payment.

Received 29 November 2002; accepted 3 February 2003; doi:10.1038/nature01474.

1. Smith, A. *The Wealth of Nations* (The Modern Library, New York, 1776).
2. Durkheim, E. *De La Division Du Travail Social* (Presses Universitaires de France, Paris, 1902).
3. Arrow, K. Gifts and exchanges. *Phil. Publ. Affairs* **1**, 343–362 (1972).
4. Knack, S. & Keefer, P. Does social capital have an economic payoff? *Q. J. Econ.* **112**, 1251–1288 (1997).
5. Bowles, S. *Economic Institutions and Behavior* (Princeton Univ. Press, Princeton, in the press).
6. Milgrom, P. & Roberts, J. *Economics, Organization & Management* (Prentice Hall, Englewood Cliffs, 1992).
7. Dawes, R. M. Social Dilemmas. *Annu. Rev. Psych.* **31**, 169–193 (1980).
8. Fehr, E., Kirchsteiger, G. & Riedl, A. Does fairness prevent market clearing? *Q. J. Econ.* **108**, 437–460 (1993).
9. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity and social history. *Games Econ. Behav.* **10**, 122–142 (1995).
10. Andreoni, J. Cooperation in public goods experiments: kindness or confusion. *Am. Econ. Rev.* **85**, 891–904 (1995).
11. Ostrom, E. A behavioral approach to the rational choice theory of collective action. *Am. Pol. Sci. Rev.* **92**, 1–22 (1998).
12. Trivers, R. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
13. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
14. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
15. Milinski, M., Semmann, D. & Krambeck, H. J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426 (2002).
16. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond. B* **268**, 745–753 (2001).
17. Zahavi, A. The cost of honesty (further remarks on the handicap principle). *J. Theor. Biol.* **67**, 603–605 (1977).
18. Mas-Colell, A., Whinston, M. D. & Green, J. R. *Microeconomic Theory* (Oxford Univ. Press, Oxford, 1995).
19. Laffont, J. J. & Martimort, D. *The Theory of Incentives* (Princeton Univ. Press, Princeton, 2002).
20. Titmuss, R. M. *The Gift Relationship* (Allen and Unwin, London, 1970).
21. Frey, B. S. *Not Just For the Money: An Economic Theory of Personal Motivation* (Edward Elgar, Cheltenham, 1997).
22. Bewley, T. F. *Why Wages Don't Fall During A Recession* (Harvard Univ. Press, Cambridge, Massachusetts, 1999).
23. Bohnet, I., Frey, B. S. & Huck, S. More order with less law: on contract enforcement, trust and crowding. *Am. Pol. Sci. Rev.* **95**, 131–144 (2001).
24. Rabin, M. Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **83**, 1281–1302 (1993).
25. Gintis, H. Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169–179 (2000).
26. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
27. Fehr, E. & Falk, A. Psychological foundations of incentives. *Eur. Econ. Rev.* **46**, 687–724 (2002).

Acknowledgements This paper is part of the European Network for the Development of Experimental Economics (ENDEAR). E.F. gratefully acknowledges support by the Ludwig Boltzmann Institut for the Analysis of Economic Growth, by the Swiss National Science Foundation and by the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms. B.R. gratefully acknowledges support by the Deutsche Forschungsgemeinschaft and the Ministerium für Wissenschaft und Forschung des Landes Nordrhein-Westfalen for providing the subject payments.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.F. (e-mail: efehr@iwe.unizh.ch).

Nanosecond radio bursts from strong plasma turbulence in the Crab pulsar

T. H. Hankins*, J. S. Kern*†, J. C. Weatherall* & J. A. Eilek*

* Physics Department, New Mexico Tech, and † National Radio Astronomy Observatory, Socorro, New Mexico 87801, USA

The Crab pulsar was discovered¹ by the occasional exceptionally bright radio pulses it emits, subsequently dubbed ‘giant’ pulses. Only two other pulsars are known to emit giant pulses^{2,3}. There is no satisfactory explanation for the occurrence of giant pulses, nor is there a complete theory of the pulsar emission mechanism in general. Competing models for the radio emission mechanism can be distinguished by the temporal structure of their coherent emission. Here we report the discovery of isolated, highly polarized, two-nanosecond sub-pulses within the giant radio pulses from the Crab pulsar. The plasma structures responsible for these emissions must be smaller than one metre in size, making them by far the smallest objects ever detected and resolved outside the Solar System, and the brightest transient radio sources in the sky. Only one of the current models—the collapse of plasma-turbulent wave packets in the pulsar magnetosphere—can account for the nanopulses we observe.

We have carried out wideband, ultrahigh time resolution observations of the Crab giant pulses at the Arecibo Observatory’s 305-m telescope. We developed a new data acquisition system, which we used at 5.5 and 8.6 GHz with bandwidths of 0.5 and 1 GHz. Data acquisition is inhibited by our system during the 30–40 s required to transfer the 8 or 16 million samples per pulse from acquisition memory to computer disk storage for offline processing. Our sample of giant pulses is therefore incomplete. However, the pulses exceeded our detection threshold typically once per minute or less, so few strong pulses were missed.

To attain nanosecond time resolution, we must fully account for the interstellar propagation effects that distort pulsar signals: dispersion, Faraday rotation and interstellar scattering broadening. The frequency-dependent delay caused by dispersive propagation through the tenuous interstellar plasma is corrected using the coherent dispersion removal technique^{4,5}. Faraday rotation is less than ten degrees across the receiver bandwidth, and scattering broadening is negligible above 5 GHz. Hence, propagation effects do not limit our time resolution.

The radio spectrum of the Crab giant pulses is steep⁶, and the radio pulses do not appear to be associated with high-energy emissions at X-ray or gamma-ray frequencies⁷. At low radio frequencies where their energies are greatest, propagation effects obliterate the detailed pulse structure. Hence, to test the detailed theoretical predictions of the emission signature, we must observe at the highest practical frequencies, where the pulses are still pristine and the lack of terrestrial radio interference permits wide receiver bandwidths.

A set of six giant pulses, which were recorded over a span of a few minutes, is shown in Fig. 1. The arrival time of a giant pulse, or the pulse phase, relative to a fixed point on the rotating star’s surface, jitters from one giant pulse to another by several hundred microseconds. However, the pulse phases are confined to the main pulse and interpulse regions of the mean profile, which is formed by averaging the intensity synchronously with the rotation period. The pulses shown here are all near the phase of the main pulse. Most of the pulses show one or more noisy bursts separated by fractions of microseconds. After coherent dedispersion, we find that the intrinsic duration of the giant pulses is extremely short; their total widths are typically less than 1 μ s above 5 GHz (interstellar scattering

broadening dominates the pulse widths below 1 GHz), and they contain clear structure down to the inverse bandwidth time limit. The intensity autocorrelation of the giant pulses shows a characteristic width (the width of the first inflection point, excluding the amplitude of the noise ‘spike’ at zero lag due to the uncorrelated receiver noise and unresolved structure) of approximately 200 ns at 5.5 and 8.6 GHz.

Pulsar radio emission has been modelled as amplitude-modulated noise^{8,9} where a wideband noise process, $n(t)$, is modulated by a more slowly varying amplitude function, $a(t)$. The noise process is produced by the ensemble of N ‘shot’ pulses, which occur randomly in the (receiver) time resolution interval, Δt . The duration of the shot pulses is related to the inverse of the pulsar spectral bandwidth, which is assumed to extend beyond $1/\Delta t$. For large N , $n(t)$ approaches a gaussian random process by the Central Limit theorem. But if the receiver resolution Δt is small enough, $N \approx 1$, the individual shots are resolved, and the observed noise process has non-gaussian statistics.

For the typical observed pulse $N \gg 1$, and the pulse is consistent with the amplitude-modulated noise model. But occasionally we record a pulse lasting tens of microseconds that is composed of extremely short, isolated, non-overlapping impulses, such as pulse 3 in Fig. 1. For these pulses $N \leq 1$, and we can distinguish individual nanopulses. In Fig. 2 we have expanded portions of pulse 3 to show the wealth of detail down to the $\Delta t = 2$ ns resolution limit imposed by the 500-MHz receiver bandwidth. We plot the right and left circular polarizations separately. It is clear that these nanopulses are strongly circularly polarized, the polarization can be of either hand, and can change hands in a few microseconds. Some giant pulses from the millisecond pulsar B1937 + 214 are known to be circularly

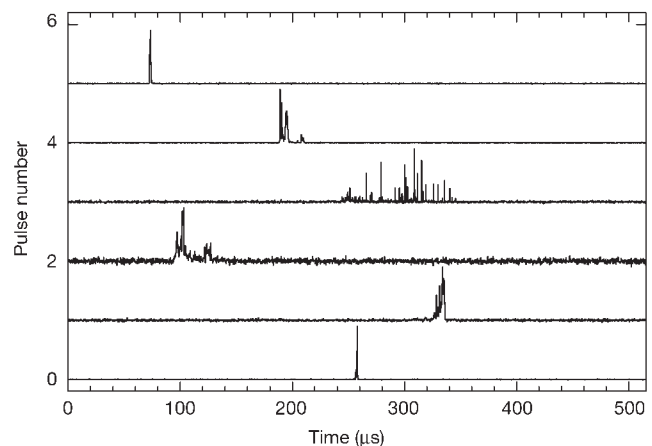


Figure 1 A sequence of dedispersed Crab giant pulses. The arrival time jitter and varied shapes of the total intensity are shown. The time axis origin is modulo one pulsar rotation period. Each pulse has been plotted with a time resolution of 250 ns and is normalized to the same maximum amplitude. The centre frequency is 5.5 GHz and the sampled bandwidth is 0.5 GHz. A square-law power detector with a 200- μ s time constant was used to detect the presence of a giant pulse in the receiver pass band. A 2-ms time window, synchronous with the Doppler-shifted main pulse arrival times, was obtained from our separate pulsar timing system. When the detected intensity exceeded a preset threshold of eight times the r.m.s. off-pulse noise during the main pulse 2-ms window, a giant pulse was captured by digitally sampling the voltage of both orthogonal polarizations at 1 or 2×10^9 samples per second using a LeCroy 9354L or LC584L digital oscilloscope. Offline, the critically sampled signals are passed through a software filter whose transfer function inverts the interstellar dispersion. The resulting time resolution is the reciprocal of the receiver bandwidth, ultimately 1–2 ns. The dedispersion process redistributes the power of any impulsive terrestrial interference over the filter impulse response time, which for these data is 1.42 ms, or 1,420,000 samples, thus reducing its impact on the pulsar signal.

polarized and can have both signs of circular polarization². Some of the nanopulses we observe are possibly shorter than 2 ns. Their intensities often exceed 10^3 Jy for extremely brief intervals, placing them only after the Sun as the brightest objects in the centimetre wavelength radio sky.

Assuming a distance to the Crab pulsar of 2.0 kpc, and an emitting source diameter $d = c\Delta t = 60$ cm, the equivalent brightness temperature of the source is 10^{37} K. This very high value for the brightness temperature implies that the emission is coherent. The energy density in a 1,000-Jy pulse of 2-ns duration is about 2×10^{14} erg cm⁻³. For comparison, standard pulsar electrodynamics¹⁰ allows us to estimate the energy density of the emitting plasma. From the charge density necessary to maintain the electric field that allows the plasma to co-rotate with the star, we obtain the number density if the plasma contains only one charge sign. If we assume the Lorentz factor of charges that initially leave the Crab pulsar is $\gamma \approx 10^7$, we obtain a plasma energy density of around 6×10^{13} erg cm⁻³. That the plasma energy density is not significantly larger than the radiation energy density clearly implies a nonlinear, collective emission process.

Statistical analysis of phase- and frequency-dependent attributes of the signals indicates that giant pulses are true temporal modulations rather than an angular beaming effect^{7,11}. If the giant pulse duration were due to the angular sweep of a beam past the observer¹², then the relativistic factor of the emitter would have to be $\gamma \approx 10^6$. Although this is comparable to the energies reached by particles initially leaving the star, in most models the radio emission is produced from a slower, denser pair plasma. Models of the pair cascade, which conserve the energy density of the initial plasma, predict much slower particle flows¹³, typically $\gamma \leq 10^3$, and therefore angular beaming is an unlikely explanation for the short time structure we observe.

If the very short nanopulses we observe are the fundamental elements of the emission, then they must result directly from the

microphysics of the radio emission process. The broad categories of processes capable of producing high brightness radio emission in pulsars—spatially coherent curvature emission¹⁴, plasma turbulence^{15,16}, and masers^{17,18}—can be tested on the basis of their behaviour on these timescales. The short timescale observations open a new window on the pulsar emission process not accessible to previous statistical studies^{19–21} that attempted to test the emission mechanism.

In direct emission mechanisms such as curvature radiation, spatial structure occurs as electrons are trapped in bunches in the potential troughs of a large-amplitude plasma wave. The emitting entity is a periodic train of bunches, of coherence length $s_0 \approx c/\Delta\nu$. For a linear column with a sinusoidal charge perturbation, the intensity of the emission is proportional to $(\langle\sigma\rangle s_0)^2$, where $\langle\sigma\rangle$ is the root-mean-square (r.m.s.) charge density of the perturbation¹³. Thus, coherent trains of short s_0 will emit less effectively than longer ones. Also, as s_0 increases, the intensity spikes grow more intense, but also narrower in frequency. On the basis of this model, Benford²² predicted that radio structure of bandwidth greater than 100 kHz would not exist, and hence one would expect no temporal features shorter than about 10 μ s. Clearly, this is in contrast to our observations. Alternatively, a reduction in coherence length of a longer ensemble of bunches could produce shorter timescale fluctuations, but a mechanism to do this has not been established.

A similar problem is inherent in maser emission models. Non-gaussian coherence effects in masers^{23,24} are possible on timescales down to the inverse of the maser bandwidth, $\Delta\nu^{-1}$. On the basis of our observed short timescale structure, $\Delta t \approx 2$ ns, a maser-like mechanism must have very wide band emission, $\Delta\nu \approx 1$ GHz. Relativistic plasma masers can interact with a broad spectrum of incoherent waves^{16,17}. However, the high brightness temperature of the transient pulses requires a long coherence path length. But, because the growth rate varies across the spectrum, the band-

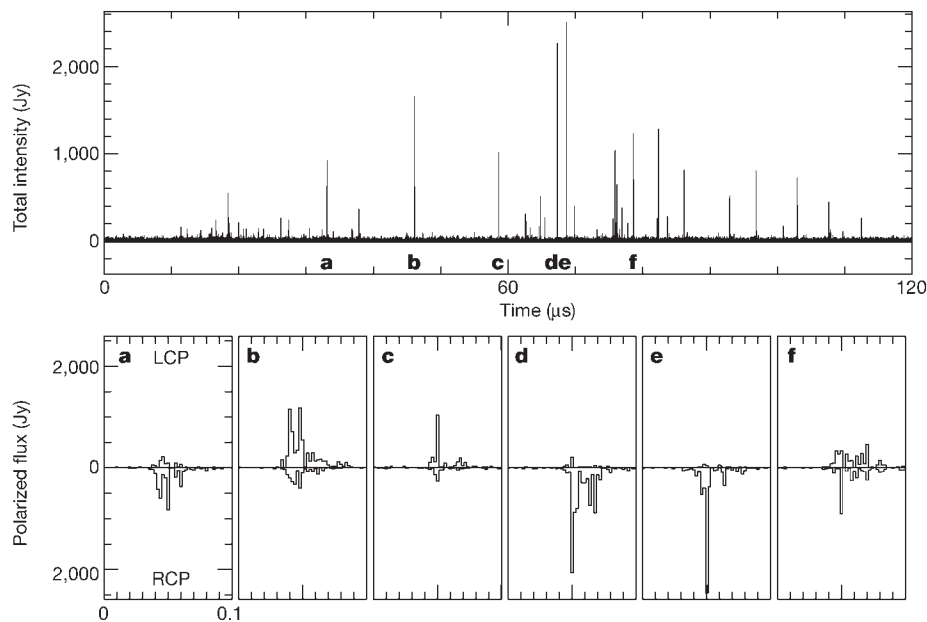


Figure 2 Intensity, polarization and time structure of nanopulses. The upper panel shows details of pulse 3 in Fig. 1, plotted with 2-ns resolution. Sections of 100-ns duration showing the polarized flux from six of the nanopulses, labelled **a–f**, are plotted below with left circular polarization upward, and right circular polarization downward, also with 2-ns resolution. The r.m.s. noise level is 18 Jy. The temporal interstellar scattering broadening, τ_{ISS} , due to multipath propagation through the turbulent Crab nebula and the interstellar

medium, is negligible; scaling our own previously measured 0.33 and 1.4 GHz values as $\nu^{-4.4}$ (Kolmogorov spectrum), we expect $\tau_{\text{ISS}} = 1.9$ ns at 5.5 GHz, and 0.2 ns at 8.6 GHz. These values are roughly consistent with interstellar scattering model predictions (J. M. Cordes and T. J. W. Lazio, personal communication) of $\tau_{\text{ISS}} = 0.5$ and 0.05 ns for $\nu = 5.5$ and 8.6 GHz, respectively.

width decreases with increasing path length. A numerical model¹⁷ for one relativistic plasma maser indicates a bandwidth that decreases to 20% for a brightness temperature of 10^{20} K. Thus, for masers, high brightness tends to be incompatible with broad bandwidth.

Emission in strong plasma turbulence is associated with the nonlinear localization of electrostatic wave energy in the plasma source region, and the subsequent explosive collapse of wave packets that takes place over a few wave oscillation periods^{25,26}. Numerical modelling²⁷ of individual pulses produced by strong plasma turbulence at 5 GHz predicts nanosecond radio bursts, and quasi-periodic structure due to resonant coupling between the plasma and radiative modes on the timescale of 1–3 ns, in agreement with our observations. Plasma turbulence is the only pulsar emission model for which a mechanism has been identified that can produce such extremely short giant radio pulses. We conclude, therefore, that giant pulse radio emission from the Crab pulsar results from the conversion of electrostatic turbulence in the pulsar magnetosphere by the mechanism of spatial collapse of nonlinear wavepackets.

The nanopulses we have detected represent an example of highly efficient coherent radio emission from a relativistic flow that serves as a prototype for coherent, transient radio emission from perhaps a larger class of sources, such as active galactic nuclei. Although extraordinary attention has been given to transients in X-ray²⁸ and gamma-ray wavelengths, radio transients represent a largely unexplored observational regime²⁹. Indeed, the techniques we apply to recover the Crab giant pulses will be useful in the search for extrasolar radio bursts and pulsars in other galaxies. □

Received 8 July 2002; accepted 4 February 2003; doi:10.1038/nature01477.

1. Staelin, D. H. & Reifenstein, E. C. III Pulsating radio sources near the Crab nebula. *Science* **162**, 1481–1483 (1968).
2. Cognard, I., Shrauner, J. A., Taylor, J. H. & Thorsett, S. E. Giant radio pulses from a millisecond pulsar. *Astrophys. J.* **457**, L81–L84 (1996).
3. Romani, R. W. & Johnston, S. Giant pulses from the millisecond pulsar B1821–24. *Astrophys. J.* **557**, L93–L96 (2001).
4. Hankins, T. H. Microsecond intensity variations in the radio emissions from CP 0950. *Astrophys. J.* **169**, 487–491 (1971).
5. Hankins, T. H. & Rickett, B. J. Pulsar signal processing. *Meth. Comp. Phys.* **14**, 55–129 (Academic, New York, 1975).
6. Sallmen, S., Backer, D. C., Hankins, T. H., Moffett, D. & Lundgren, S. Simultaneous dual-frequency observations of giant pulses from the Crab pulsar. *Astrophys. J.* **517**, 460–471 (1999).
7. Lundgren, S. C. *et al.* Giant pulses from the Crab pulsar: A joint radio and gamma-ray study. *Astrophys. J.* **453**, 433–446 (1995).
8. Rickett, B. J. Amplitude-modulated noise: An empirical model for the radio radiation received from pulsars. *Astrophys. J.* **197**, 185–191 (1975).
9. Cordes, J. M. Pulsar radiation as polarized shot noise. *Astrophys. J.* **210**, 780–791 (1976).
10. Goldreich, P. & Julian, W. H. Pulsar electrodynamics. *Astrophys. J.* **157**, 869–880 (1969).
11. Rickett, B. J. & Cordes, J. M. In *Pulsars, IAU Symp. 95* (eds Sieber, W. & Wielebinski, R.) 107–109 (Reidel, Boston, 1981).
12. Popov, M. V. *et al.* Microstructure of pulsar radio pulses measured with a time resolution of 62.5 ns at 1650 MHz. *Astron. Rep.* **46**, 206–213 (2002).
13. Arendt, P. N. Jr & Eilek, J. A. Pair creation in the pulsar magnetosphere. *Astrophys. J.* **581** (in the press).
14. Buschauer, R. & Benford, G. General theory of coherent curvature radiation. *Mon. Not. R. Astron. Soc.* **177**, 109–136 (1976).
15. Asseo, E., Pelletier, G. & Sol, H. A non-linear radio pulsar emission mechanism. *Mon. Not. R. Astron. Soc.* **247**, 529–548 (1990).
16. Weatherall, J. C. Modulational instability, mode conversion, and radio emission in the magnetized pair plasma of pulsars. *Astrophys. J.* **483**, 402–413 (1997).
17. Lyutikov, M., Blandford, R. D. & Machabeli, G. On the nature of pulsar radio emission. *Mon. Not. R. Astron. Soc.* **305**, 338–352 (1999).
18. Weatherall, J. C. A relativistic-plasma Compton maser. *Astrophys. J.* **559**, 196–200 (2001).
19. Jenet, F. A., Anderson, S. B. & Prince, T. A. The first detection of coherent emission from radio pulsars. *Astrophys. J.* **558**, 302–308 (2001).
20. Cairns, I. H., Johnston, S. & Das, P. Intrinsic variability of the Vela pulsar: Lognormal statistics and theoretical implications. *Astrophys. J.* **563**, L65–L68 (2001).
21. Delaney, T. & Weatherall, J. C. Model for deterministic chaos in pulsar radio signals and search for attractors in the Crab and Vela pulsars. *Astrophys. J.* **519**, 291–302 (1999).
22. Benford, G. Model for the microstructure emission of pulsars. *Mon. Not. R. Astron. Soc.* **179**, 311–315 (1977).
23. Elitzur, M. *Astronomical Masers* 68–71 (Kluwer, Boston, 1992).
24. Moran, J. M. Statistical properties of the radiation fields from H₂O masers. *Bull. Am. Astron. Soc.* **13**, 508 (1981).
25. Zakharov, V. E. Collapse of Langmuir waves. *Sov. Phys. JETP* **35**, 908–914 (1972).
26. Goldman, M. V. Strong turbulence of plasma waves. *Rev. Mod. Phys.* **56**, 709–735 (1984).

27. Weatherall, J. C. Pulsar radio emission by conversion of plasma wave turbulence: Nanosecond time structure. *Astrophys. J.* **506**, 341–346 (1998).
28. Strohmayer, T. E., Swank, J. H. & Zhang, W. in *The Active X-ray Sky: Results from BeppoSAX and RXTE* (eds Scarsi, L., Bradt, H., Giommi, P. & Fiore, F.) *Nucl. Phys. B (Proc. Suppl.)* **69**, 129 (1997).
29. Mészáros, P. Gamma-ray bursts: Accumulating afterglow implications, progenitor clues, and prospects. *Science* **291**, 79–84 (2001).

Acknowledgements T.H.H. thanks NRAO and NAIC for partial sabbatical leave support at the Green Bank and Arecibo Observatories. J.S.K. thanks NRAO for pre-doctoral support. We thank B. Driggers of TMS and LeCroy for the loan of a LeCroy LT584L oscilloscope, J. Ford at NRAO Green Bank for square-law detectors, NRAO Socorro for computer support and technical assistance, and the NSF for a research grant. The Arecibo Observatory is operated by Cornell University under a cooperative agreement with the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.H.H. (e-mail: thankins@aoc.nrao.edu).

An extended upper atmosphere around the extrasolar planet HD209458b

A. Vidal-Madjar*, A. Lecavelier des Etangs*, J.-M. Désert*, G. E. Ballester†, R. Ferlet*, G. Hébrard* & M. Mayor‡

* Institut d'Astrophysique de Paris, CNRS/UPMC, 98bis boulevard Arago, F-75014 Paris, France

† Lunar and Planetary Laboratory, University of Arizona, 1040 E. 4th St., Rm 901, Tucson, Arizona 85721-0077, USA

‡ Observatoire de Genève, CH-1290 Sauverny, Switzerland

The planet in the system HD209458 is the first one for which repeated transits across the stellar disk have been observed^{1,2}. Together with radial velocity measurements³, this has led to a determination of the planet's radius and mass, confirming it to be a gas giant. But despite numerous searches for an atmospheric signature^{4–6}, only the dense lower atmosphere of HD209458b has been observed, through the detection of neutral sodium absorption⁷. Here we report the detection of atomic hydrogen absorption in the stellar Lyman α line during three transits of HD209458b. An absorption of $15 \pm 4\%$ (1σ) is observed. Comparison with models shows that this absorption should take place beyond the Roche limit and therefore can be understood in terms of escaping hydrogen atoms.

Far more abundant than any other species, hydrogen is well-suited for searching weak atmospheric absorptions during the transit of an extrasolar giant planet in front of its parent star, in particular over the strong resonant stellar ultraviolet Lyman α emission line at 1,215.67 Å. Depending upon the characteristics of the planet's upper atmosphere, an H I signature much larger than that for Na I at 0.02% (ref. 7) is foreseeable. Three transits of HD209458b (named A, B and C hereafter) were sampled in 2001 (on 7–8 September, 14–15 September and 20 October, respectively) with the Space Telescope Imaging Spectrograph (STIS) onboard the Hubble Space Telescope (HST); the data set is now public in the HST archive. To partially overcome contamination from the Earth's Lyman α geocoronal emission, we used the G140M grating with the $52'' \times 0.1''$ slit (medium spectral resolution: $\sim 20 \text{ km s}^{-1}$). For each transit, three consecutive HST orbits (named 1, 2 and 3 hereafter) were scheduled such that the first orbit (1,780 s exposure) ended before the first contact to serve as a reference, and the two following

ones (2,100 s exposures each) were partly or entirely within the transit.

The observed Lyman α spectrum of HD209458 is typical for a solar-type star, with a double-peaked emission originating from the stellar chromosphere (Fig. 1). It also shows a wide central absorption feature due to neutral hydrogen in the interstellar medium. The geocoronal emission filled the aperture of the spectrograph, resulting in an extended emission line perpendicular to the dispersion direction. The extent of this emission along the slit allowed us to remove it at the position of the target star. We evaluated its variation both along the slit and from one exposure to another, and excluded the wavelength domain where the corresponding standard deviation per pixel is larger than 20% of the final spectrum. We concluded that the geocoronal contamination can be removed with high enough accuracy outside the central region ($1,215.5 \text{ \AA} < \lambda < 1,215.8 \text{ \AA}$, labelled 'Geo' in Figs 2, 4).

From the two-dimensional (2D) images of the far-ultraviolet (FUV) multi-anode microchannel array (MAMA) detector, the STIS standard pipeline extracts one-dimensional spectra in which the dark background has not been removed. The background level was systematically increasing from one exposure to the next within each of the three visits, but still remained below 2% of the peak intensity of the stellar signal. We therefore reprocessed the 2D images by using two independent approaches. The first one uses the 2D images provided by the standard pipeline and interpolates the background (including the Earth's geocoronal emission) with a polynomial fitted per column above and below the spectrum region. The second method starts from the 2D raw images to which is applied a dark background correction from a super-dark image (created for the period of the observations) as well as a subtraction of the geocoronal emission measured along the slit, away from the stellar spectrum. The differences between the results of both approaches are negligible, showing that systematic errors generated through the background corrections are small compared to the statistical errors. Those errors are dominated by photon counting

noise to which we added quadratically the error evaluated for both the dark background and geocoronal subtractions.

The Lyman α line profiles observed before and during the transits are plotted in Fig. 2. The three exposures outside the transits (exposures A1, B1 and C1) and the three entirely within the transits (A2, B3 and C3) were co-added to improve the signal-to-noise ratio. An obvious signature in absorption is detected during the transits, mainly over the blue side of the line, and possibly at the top of the red peak.

To characterize this signature better, we have defined two spectral domains: 'In' and 'Out' of the absorption. The In domain is a wavelength interval limited by two variables λ_1 and λ_2 (excluding the Geo geocoronal region). The Out domain is the remaining wavelength coverage within the interval $1,214.4\text{--}1,216.8 \text{ \AA}$, for which the Lyman α line intensity can be accurately measured at the time of the observation. The corresponding In/Out flux ratio derived for each exposure is shown in Fig. 3, revealing the absorption occurring during the transits. To evaluate whether this detection is sensitive to a particular choice of λ_1 and λ_2 , we averaged the three ratios before the transits and the three ratios entirely within the transits. We calculated the averaged pre-transits over mid-transits ratio as a function of λ_1 and λ_2 and propagated the errors through boot-strap estimations of the ratio calculated with 10,000 randomly generated spectra according to the evaluated errors over each individual pixel (Fig. 2c). The averaged ratio is always significantly below 1, with the minimum at $\lambda_1 = 1,215.15 \text{ \AA}$ and $\lambda_2 = 1,216.1 \text{ \AA}$. In the interval defined by these two wavelengths, the Lyman α line is reduced by $15 \pm 4\%$ (1σ) during the transit. This is a larger-than- 3σ detection of an absorption in the hydrogen line profile during the planetary transits.

HD209458 (G0V) is close to solar type, for which time variations are known to occur in the chromospheric Lyman α line⁸. We thus evaluated the In/Out ratio in the solar Lyman α line profile as measured over the whole solar disk by the Solar Ultraviolet Measurements of Emitted Radiation (SUMER) instrument onboard

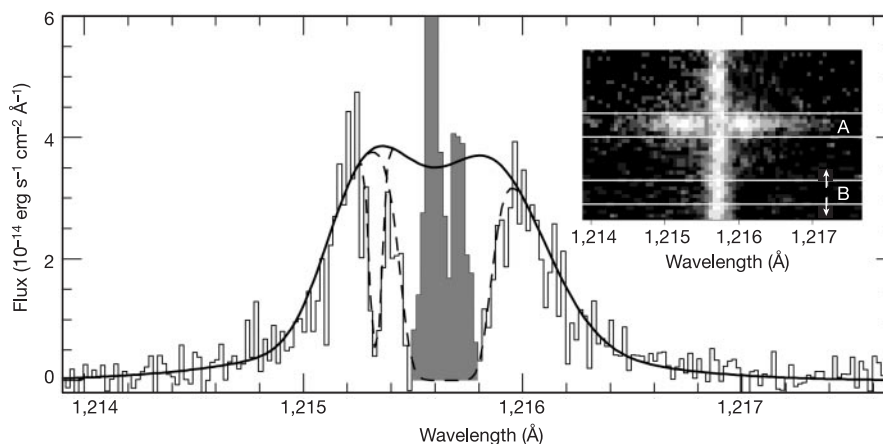


Figure 1 The HD209458 Lyman α emission line. This high-resolution spectrum (histogram) was obtained with the E140M echelle grating and the $0.2'' \times 0.2''$ wide slit with a spectral resolution of 5 km s^{-1} ; it was not used in the present analysis, but it allows the different components of the line profile to be seen. The continuum is a double-peaked emission line originating from the stellar chromosphere: the temperature increase in the lower chromosphere causes an emission line with a central dip due to the high opacity of the abundant hydrogen atoms (solid line). The observed spectrum also has a narrow absorption line ($1,215.3 \text{ \AA}$, barely seen at lower resolution) and a central wide absorption line ($1,215.6 \text{ \AA}$) due to the interstellar deuterium and hydrogen, respectively (dashed lines). The grey zone represents the fraction of the spectrum contaminated by the geocoronal emission, which is double-peaked in that case because the plotted spectrum

is the average of four exposures obtained at two different epochs. The inset shows a small portion of the 2D image of a G140M first-order spectrum containing the stellar Lyman α profile and a sample of the geocoronal signal. This spectrum is one of the nine spectra used for this analysis. The G140M spectra have lower spectral resolution but higher signal-to-noise ratio. The stellar spectrum is seen as a horizontal line where the two peaks are resolved from the geocoronal emission (vertical line along the slit). The one-dimensional spectra are obtained by vertically adding around ten pixels within the A band. Measurements along the slit direction (~ 800 pixels available), for example at the position of the B band, allow us to estimate the geocoronal contamination and the background subtraction as well as the corresponding uncertainties.

the Solar and Heliospheric Observatory (SOHO)⁹ during the last solar cycle from 1996 to 2001, that is, from quiet to active Sun. During this time, the total solar Lyman α flux varies by about a factor of two, while its In/Out ratio varies by less than $\pm 6\%$. Within a few months, a time comparable to our HD209458 observations, the solar In/Out ratio varies by less than $\pm 4\%$. This is an indication that the absorption detected is not of stellar origin but is due to a transient absorption occurring during the planetary transits.

A bright hot spot on the stellar surface hidden during the planetary transit is also excluded. Such a hot spot would have to contribute about 15% of the Lyman α flux over 1.5% of the stellar surface occulted by the planet, in contradiction with Lyman α inhomogeneities observed on the Sun¹⁰. Furthermore, this spot would have to be perfectly aligned with the planet throughout the transit, at the same latitude as the Earth's direction, and with a peculiar narrow single-peaked profile confined over the In spectral region. It seems unlikely that a stellar spot could satisfy all these conditions.

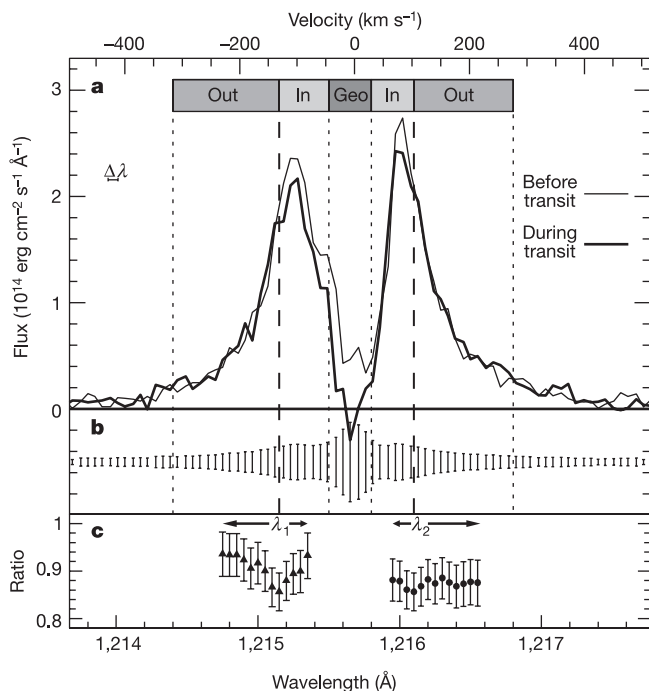


Figure 2 The HD209458 Lyman α profile observed with the G140M grating. The geocoronal emission has been subtracted; the propagated errors are consequently larger in the central part of the profile, particularly in the Geo domain (see text). $\Delta\lambda$ represents the spectral resolution. **a**, The thin line shows the average of the three observations performed before the transits (exposures A1, B1 and C1); the thick line shows the average of the three observations recorded entirely within the transits (exposures A2, B3 and C3). Variations are seen in the In domain as absorption over the blue peak of the line and partially over the red peak (between -130 km s^{-1} and 100 km s^{-1}). Quoted velocities are in the stellar reference frame, centred on -13 km s^{-1} in the heliocentric reference frame. **b**, $\pm 1\sigma$ error bars. **c**, The ratio of the two spectra in the In domain, the spectra being normalized such that the ratio is 1 in the Out domain. This ratio is plotted as a function of λ_1 using $\lambda_2 = 1,216.10 \text{ \AA}$ (triangles), and as a function of λ_2 using $\lambda_1 = 1,215.15 \text{ \AA}$ (circles). The ratio is always significantly below 1, with a minimum at $\lambda_1 = 1,215.15 \text{ \AA}$ (-130 km s^{-1}) and $\lambda_2 = 1,216.10 \text{ \AA}$ (100 km s^{-1}). In the domain defined by these values, the Lyman α intensity decreases during the transits by $15 \pm 4\%$. The detection does not strongly depend on a particular selection of the domain. While the decrease of the Lyman α intensity is not sensitive to the position of λ_2 , it is more sensitive to the position of λ_1 , showing that most of the absorption occurs in the blue part of the line. Using the whole domain where the absorption is detected, the exoplanetary atmospheric hydrogen is detected at more than 3σ .

Finally, we confirmed with various tests that there are no correlations between the geocoronal variations and the detected signature in absorption. One method is presented in Fig. 4, showing that a contamination of the In domain by the geocorona is excluded. We thus conclude that the detected profile variation can only be related to an absorption produced by the planetary environment.

The observed 15% intensity drop is larger than expected *a priori* for an atmosphere of a planet occulting only 1.5% of the star. Although the small distance (8.5 stellar radii) between the planet and the star results in an extended Roche lobe¹¹ with a limit at about 2.7 planetary radii (that is, 3.6 Jupiter radii), the filling up of this lobe gives a maximum absorption of about 10% during the planetary transit. Because a more important absorption is detected, hydrogen atoms must cover a larger area: a drop of 15% corresponds to an occultation by an object of 4.3 Jupiter radii. This is clearly beyond the Roche limit as theoretically predicted⁶. Thus some hydrogen atoms should escape from the planet. The spectral absorption width shows independently that the atoms have large velocities relative to the planet. Thus hydrogen atoms must be escaping the planetary atmosphere.

We have built a particle simulation in which we assumed that hydrogen atoms are sensitive to the stellar radiation pressure

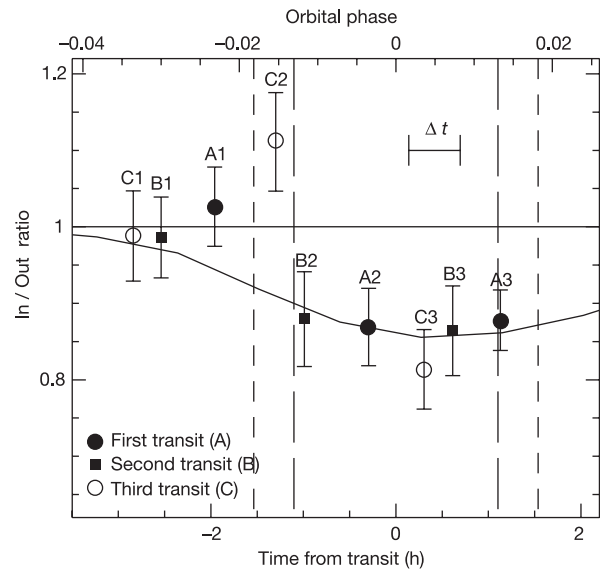


Figure 3 Relative flux of Lyman α as a function of the HD209458's system phase. The averaged ratio of the flux is measured in the In ($1,215.15\text{--}1,215.50 \text{ \AA}$ and $1,215.80\text{--}1,216.10 \text{ \AA}$) and the Out ($1,214.40\text{--}1,215.15 \text{ \AA}$ and $1,216.10\text{--}1,216.80 \text{ \AA}$) domains in individual exposures of the three observed transits of HD209458b. The central time of each exposure is plotted relative to the transit time. The vertical dashed lines indicate the first and the second contact at the beginning and the end of the transit; the exposures A1, B1 and C1 were performed before the transits, and the exposures A2, B3 and C3 were entirely within the transits. The ratio is normalized to the average value of the three observations completed before the beginning of the transits. The $\pm 1\sigma$ error bars are statistical; they are computed through boot-strap estimations (see text). The In/Out ratio smoothly decreases by around 15% during the transit. The thick line represents the absorption ratio modelled through a particle simulation which includes hydrogen atoms escaping from the planet. In this simulation, hydrogen atoms are sensitive to the radiation pressure above an altitude of 0.5 times the Roche radius, where the density is assumed to be $2 \times 10^5 \text{ cm}^{-3}$; these two parameters correspond to an escape flux of $\sim 10^{10} \text{ g s}^{-1}$. The stellar radiation pressure is taken to be 0.7 times the stellar gravitation. The mean lifetime of escaping hydrogen atoms is taken to be 4 h. The model yields an atom population in a curved comet-like tail, explaining why the computed absorption lasts well after the end of the transit.

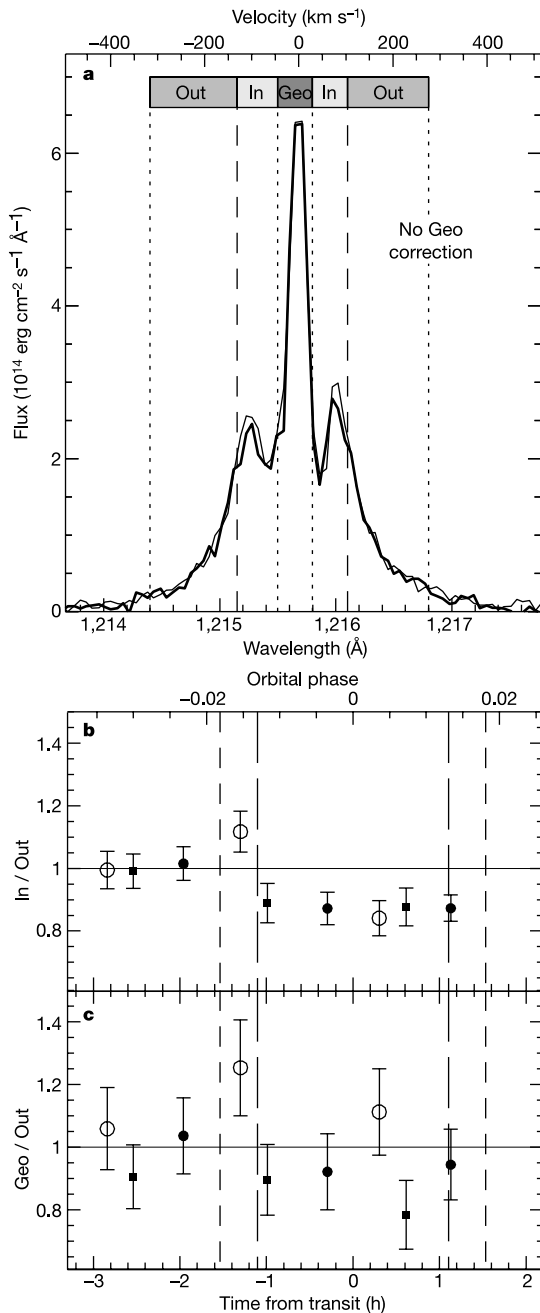


Figure 4 Spectra and ratios with no geocoronal correction. To investigate the possibility that a bad estimate of the geocoronal correction may cause the detected signal, the spectra and ratios are plotted without this correction. First, note that the geocoronal emission is of the order of the stellar flux, and its peak value is never larger than three times the stellar one. Consequently, the correction is very small outside the Geo domain. **b** shows the same evaluation of the In/Out ratios as made in Fig. 3, with no geocoronal and no background correction. It appears that the impact of the geocoronal correction on these ratios is negligible. To further show that the ‘wings’ of the geocoronal emission do not carry the detected transit signal, **c** shows the Geo/Out ratios, where Geo is the total flux due to the geocorona. If the flux in the In domain was significantly affected by the geocorona, the Geo/Out ratios should present a signal similar to the In/Out ratios plotted in Fig. 3. The Geo/Out ratios are found to be random fluctuations around the average of the A1, B1 and C1 values. There is no consistent signature of the transit in the geocoronal variations. As in Fig. 2a, **a** shows the average of the three spectra obtained before and during the transits (thin and thick lines, respectively). Quoted velocities are in the stellar reference frame. Here the spectra are without any geocoronal correction. A simple constant has been subtracted to compensate for the mean background levels in order to have matching spectra in the Out domain. This clearly shows that the absorption signature is present even without the geocoronal correction.

inside and outside the Roche lobe. Their motion is evaluated by taking into account both the planetary and stellar gravities. The Lyman α radiation pressure is known to be 0.7 times the stellar gravity, the escape flux and the neutral hydrogen lifetime being free parameters. The lifetime of hydrogen is limited to a few hours owing to stellar extreme ultraviolet ionization. Escaping hydrogen atoms expand in an asymmetric comet-like tail and progressively disappear when moving away from the planet. This simple scenario is consistent with the observations (Fig. 3). In this model, atoms in the evaporating coma and tail cover a large area of the star, and most are blueshifted because of the radiation pressure repelling them away from the star. The detection of most of the absorption in the blue part of the line is consistent with these escaping atoms. On the other hand, more observations are needed to clarify whether an absorption is also present in the red part of the line.

To account for the observed absorption depth, the particle simulation implies a minimum escape flux of around 10^{10} g s⁻¹. However, owing to saturation effects in the absorption line, a flux larger by several orders of magnitude would produce a similar absorption signature. So, to evaluate the actual escape flux, we need to estimate the vertical distribution of hydrogen atoms up to the Roche limit, in an atmosphere extended by the stellar tidal forces and heated by many possible mechanisms. These effects may lead to a much larger escape flux. A detailed calculation is beyond the scope of this Letter. This raises the question of the lifetime of evaporating extrasolar planets which may be comparable to the star’s lifetime itself. If so, the so-called ‘hot Jupiters’ could evolve faster than their parent star, eventually becoming smaller objects, which could look like ‘hot hydrogen-poor Neptune-mass’ planets. This evaporation process, more efficient for planets close to their star, might explain the very few detections^{12,13} of ‘hot Jupiters’ with orbiting periods shorter than three days. □

Received 13 September 2002; accepted 27 January 2003; doi:10.1038/nature01448.

1. Henry, G. W., Marcy, G. W., Butler, R. P. & Vogt, S. S. A transiting “51 Peg-like” planet. *Astrophys. J.* **529**, L41–L44 (2000).
2. Charbonneau, D., Brown, T. M., Latham, D. W. & Mayor, M. Detection of planetary transits across a Sun-like star. *Astrophys. J.* **529**, L45–L48 (2000).
3. Mazeh, T. et al. The spectroscopic orbit of the planetary companion transiting HD 209458. *Astrophys. J.* **532**, L55–L58 (2000).
4. Bundy, K. A. & Marcy, G. W. A search for transit effects in spectra of 51 Pegasi and HD 209458. *Proc. Astron. Soc. Pacif.* **112**, 1421–1425 (2000).
5. Rauer, H., Bockelée-Morvan, D., Coustenis, A., Guillot, T. & Schneider, J. Search for an exosphere around 51 Pegasi B with ISO. *Astron. Astrophys.* **355**, 573–580 (2000).
6. Moutou, C. et al. Search for spectroscopical signatures of transiting HD 209458b’s exosphere. *Astron. Astrophys.* **371**, 260–266 (2001).
7. Charbonneau, D., Brown, T. M., Noyes, R. W. & Gilliland, R. L. Detection of an extrasolar planet atmosphere. *Astrophys. J.* **568**, 377–384 (2002).
8. Vidal-Madjar, A. Evolution of the solar Lyman α flux during four consecutive years. *Sol. Phys.* **40**, 69–86 (1975).
9. Lemaire, P. et al. in *Proc. Symp. SOHO 11, From Solar Min to Max: Half a Solar Cycle with SOHO* (ed. Wilson, A.) 219–222 (ESA SP-508, ESA Publications Division, Noordwijk, 2002).
10. Prinz, D. K. The spatial distribution of Lyman α on the Sun. *Astrophys. J.* **187**, 369–375 (1974).
11. Paczynski, B. Evolutionary processes in close binary systems. *Annu. Rev. Astron. Astrophys.* **9**, 183–208 (1971).
12. Cumming, A., Marcy, G. W., Butler, R. P. & Vogt, S. S. The statistics of extrasolar planets: Results from the Keck survey. Preprint astro-ph/0209199 available at (<http://xxx.lanl.gov>) (2002).
13. Konacki, M., Torres, G., Jha, S. & Sasselov, D. D. An extrasolar planet that transits the disk of its parent star. *Nature* **421**, 507–509 (2003).

Acknowledgements This work is based on observations with the NASA/ESA Hubble Space Telescope, obtained at the Space Telescope Science Institute, which is operated by AURA, Inc. We thank M. Lemoine, L. Ben Jaffel, C. Emerich, P. D. Feldman and J. McConnell for comments, J. Herbert and W. Landsman for conversations on STIS data reduction, and J. Valenti for help in preparing the observations.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.V.-M. (e-mail: alfred@iap.fr).

Observation of two-dimensional discrete solitons in optically induced nonlinear photonic lattices

Jason W. Fleischer^{*†}, Mordechai Segev^{*†}, Nikolaos K. Efremidis[‡] & Demetrios N. Christodoulides[‡]

^{*} Physics Department, Technion—Israel Institute of Technology, Haifa 32000, Israel

[†] Electrical Engineering Department, Princeton University, New Jersey 08544, USA

[‡] School of Optics/CREOL, University of Central Florida, Florida 32816-2700, USA

Nonlinear periodic lattices occur in a large variety of systems, such as biological molecules¹, nonlinear optical waveguides², solid-state systems³ and Bose–Einstein condensates⁴. The underlying dynamics in these systems is dominated by the interplay between tunnelling between adjacent potential wells and nonlinearity^{1–15}. A balance between these two effects can result in a self-localized state: a lattice or ‘discrete’ soliton^{1,2}. Direct observation of lattice solitons has so far been limited to one-dimensional systems, namely in arrays of nonlinear optical waveguides^{2,9–17}. However, many fundamental features are expected to occur in higher dimensions, such as vortex lattice solitons¹⁸, bright lattice solitons that carry angular momentum, and three-dimensional collisions between lattice solitons. Here, we report the experimental observation of two-dimensional (2D) lattice solitons. We use optical induction, the interference of two or more plane waves in a photosensitive material, to create a 2D photonic lattice in which the solitons form^{11,12}. Our results pave the way for the realization of a variety of nonlinear localization phenomena in photonic lattices and crystals^{19–23}. Finally, our observation directly relates to the proposed lattice solitons in Bose–Einstein condensates⁴, which can be observed in optically induced periodic potentials^{24,25}.

In general, wave propagation in periodic lattices (such as an array of optical waveguides) is fundamentally different from that occurring in a homogeneous medium. For example, when light is focused into one waveguide, linear propagation along the waveguides results in tunnelling to adjacent sites, exhibiting a characteristic diffraction pattern with the intensity mainly concentrated in the outer lobes. For a sufficiently high nonlinearity, self-focusing

can balance this effect, leading to a lattice (discrete) soliton^{2,9,10}. For light propagating at an angle $\theta = k_x/k \approx k_x/k_z$ with respect to the array, the periodicity of the lattice becomes important, as the corresponding ‘Bloch momentum’ k_x can satisfy Bragg reflection conditions within the Brillouin zone (defined in the range $|k_x D| \leq \pi$, where D is the lattice spacing). Near the edge of this zone, diffraction becomes anomalous (‘negative’), leading to such effects as diffraction management^{13,14} and staggered (π out-of-phase) solitons^{11,15,17}. These are some of the fundamental aspects of solitons in nonlinear periodic structures, which were originally discovered through the theoretical paradigm of the discrete nonlinear Schrödinger equation^{1–4,16}, hence the term ‘discrete soliton’^{1,2}. Experimentally, self-localized ‘breathers’ have been observed in various physical settings^{5–8}, but lattice (discrete) solitons have thus far been reported only in nonlinear optical systems and only in one-dimensional (1D) configurations^{9–11,13,17}. In what follows we demonstrate bright 2D lattice solitons in their simplest realization: in-phase solitons at the base of the first Brillouin zone. In addition, we demonstrate bright self-trapped wave packets at the edge of the first Brillouin zone.

The formation of the 2D nonlinear photonic lattice relies on an optical induction technique^{11,12} in which a 2D array of waveguides is induced in a nonlinear medium. We proposed this method theoretically¹² and recently demonstrated experimentally¹¹ 1D lattice solitons in a 1D waveguide array. The waveguide array is induced, in real time, in a photosensitive material by interfering two or more plane waves. A separate ‘probe’ beam is launched into the periodic waveguide array, where it exhibits discrete diffraction and, at a sufficiently high nonlinearity, forms a lattice soliton. For this system to work, it is essential that the waveguides are as uniform as possible, implying that the interference pattern (inducing the lattice) must not change in the propagation direction. For this to happen in the nonlinear medium, the interfering waves themselves should not be affected by the nonlinearity. At the same time, the probe (soliton-forming) beam must experience the highest possible nonlinearity. A photorefractive material with a strong electro-optic anisotropy allows this scenario; the interfering beams are polarized in a non-electro-optic direction and the probe is polarized along the crystal-line c axis. In this arrangement, the interfering beams will propagate mostly linearly, while the signal beam will experience both a periodic potential and a significant (screening) nonlinearity^{26,27}. In this way we have demonstrated both on-axis and staggered 1D lattice solitons¹¹. However, 1D systems, although quite instructive, cannot host a variety of fascinating nonlinear phenomena that require a higher lattice dimensionality. Our method of optical lattice induction allows for dynamic, reconfigurable arrays of almost any geometry.

This method of optical induction is quite general, allowing the

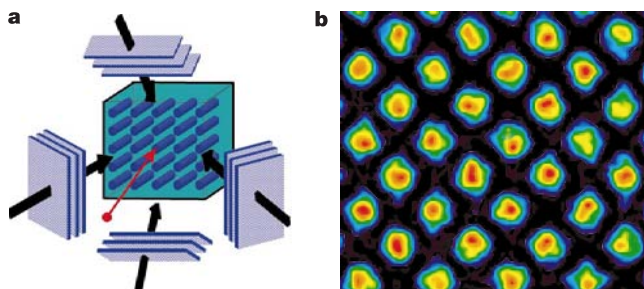


Figure 1 Experimental scheme and a typical photonic lattice. **a**, Diagram of our experimental set-up. We use a photosensitive (photorefractive) crystal with electro-optic anisotropy: two interfering pairs of ordinarily polarized plane waves induce the photonic array, while the extraordinarily polarized probe (soliton-forming) beam is focused into a single waveguide. **b**, Typical observation of a waveguide array at the exit face of the crystal. Each waveguide is approximately 7 μm in diameter, with an 11 μm spacing between nearest neighbours.

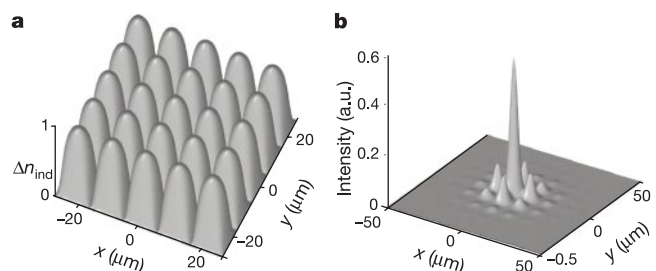


Figure 2 Numerical simulation results depicting the induced photonic lattice and the structure of an on-axis lattice soliton. **a**, Calculated 2D structure of the induced index change (photonic lattice) for the self-focusing nonlinearity. **b**, Simulated intensity structure of the on-axis lattice soliton.

simple creation of photonic lattices in any photosensitive medium of sufficient anisotropy. Theoretically, the dynamical evolution of the system along z is governed by two coupled equations that describe the slowly-varying amplitudes of the lattice wave V (periodic along x and y) and the (soliton-forming) probe U (ref. 12):

$$i\frac{\partial U}{\partial z} + \frac{1}{2k_1}\left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}\right) - \Delta n_1(I)U = 0 \quad (1)$$

$$i\frac{\partial V}{\partial z} + \frac{1}{2k_2}\left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2}\right) - \Delta n_2(I)V = 0 \quad (2)$$

Here, the subscripts indicate the orthogonal polarization directions, n_1 and n_2 are the respective indices of refraction, $k_1 = k_0 n_1$, $k_2 = k_0 n_2$, and Δn_1 , Δn_2 are the nonlinear index changes induced by the total intensity $I = |U|^2 + |V|^2$. For our specific experiment, $\Delta n_1 = [k_0 n_e^3 r_{33} E_0 / 2][1 + I]^{-1}$ and $\Delta n_2 = [k_0 n_o^3 r_{13} E_0 / 2][1 + I]^{-1}$ are the index changes created by the photorefractive screening nonlinearity^{26,27}, where the intensity I is measured in units of the background illumination. In these expressions, E_0 is the applied field and $\{n_1 = n_e, r_{33}\}$ and $\{n_2 = n_o, r_{13}\}$ are the indices of refraction and electro-optic coefficient for the extraordinarily polarized probe beam U and ordinarily polarized array wave V , respectively. We chose a highly anisotropic SBN:75 crystal, in which the nonlinear coefficient $n_o^3 r_{13}$ for V is more than 20 times smaller than the nonlinear coefficient $n_e^3 r_{33}$ for U , implying that Δn_2 is negligible and the dynamics of the array wave is indeed linear. (Note that under suitable conditions, this reduced system of equations can further simplify¹² into a discrete nonlinear Schrödinger equation².) An additional assumption in equation (1) is that the photorefractive screening nonlinearity acts isotropically in x and y . This is well justified for our parameters considering that photorefractive screening solitons are almost circular, as are their induced waveguides²⁸. The periodic structure of the waveguide array is implicitly contained in the potential induced by $|V(x, y)|^2$. For our representative case of a square lattice of side D formed by four symmetrically interfering plane waves, the lattice intensity $|V|^2 = V_0^2 [\cos(\pi x/D) + \cos(\pi y/D)]^2$ provides a periodic potential for the probe field U . In this scenario, the probe (soliton) beam experiences the familiar competition between nearest-neighbour coupling and nonlinearity. We note that the sign of our nonlinearity (whether it is self-focusing or self-defocusing) as well as its magnitude, can be controlled through the polarity and strength of the bias field E_0 .

Our experimental set-up is sketched in Fig. 1a. We use a 6-mm-long SBN:75 crystal, with electro-optic coefficients $r_{13} \approx 67 \text{ pm V}^{-1}$ and $r_{33} \approx 1,340 \text{ pm V}^{-1}$. As a representative geometry, we induce a 2D square photonic lattice (2D array of 2D waveguides) by interfering two pairs of ordinarily polarized plane waves. A typical experimental result (extracted from a $6 \text{ mm} \times 6 \text{ mm}$ optically induced photonic lattice) is shown in Fig. 1b, in which each waveguide has a diameter of about $7 \mu\text{m}$, with an $11 \mu\text{m}$ separation between its nearest neighbours. The probe (soliton-forming) beam is extraordinarily polarized and is focused into one of the waveguides, both to maximize the propagation effects through the periodic potential and to yield the narrowest possible lattice solitons. (We note that we are able to generate broader solitons by launching the probe beam into more than one channel and using lower applied fields.) Each array-forming plane wave has 15 mW of power, while the intensity ratio between the entire optical lattice and the probe beam is 5:1. Voltage applied against (along) the c axis sets the focusing (or defocusing) photorefractive screening nonlinearity and, in the proper parameter range, leads to localization of the probe beam and to the formation of lattice solitons. The crystal is also illuminated uniformly (from the top) with a background beam of white light, facilitating fine-tuning of the saturation level of the photorefractive screening nonlinearity.

We first consider on-axis propagation and in-phase lattice solitons. In this regime, the probe beam U experiences normal diffraction, so bright solitons necessitate a self-focusing nonlinearity for which E_0 must be positive. The calculated 2D structure of the induced index change (photonic lattice) and the intensity of the on-axis soliton are shown in Figs 2a and b, respectively. Our experimental results are shown in Fig. 3. Shown is the intensity of the probe beam as it exits the photonic lattice after propagating 6 mm along the 2D waveguide array. This probe beam is launched into a single (the central) waveguide. At low voltages, the probe propagates linearly, and as a result discrete diffraction concentrates the signal intensity into the outer perimeter of a square (Fig. 3a). For a stronger nonlinearity (at a higher voltage), self-focusing dominates and a lattice soliton (a 2D discrete soliton) forms (Fig. 3b). The soliton structure consists of a central intensity peak surrounded by weaker side lobes. An interferogram of this soliton, obtained by interfering the soliton output beam with a plane wave, shows constructive interference of all the elements; that is, the central peak is in-phase with its neighbours (Fig. 3c). As the signal intensity

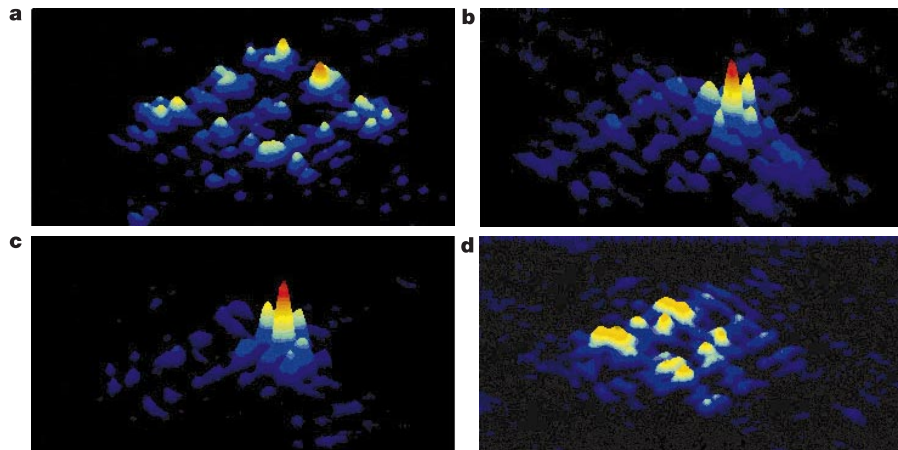


Figure 3 Experimental results presenting the propagation of a probe beam launched into a single waveguide at normal incidence (on-axis propagation). **a**, Intensity structure of the probe beam at the exit face of the crystal, displaying discrete diffraction at low nonlinearity (200 V). **b**, Intensity structure of the probe beam at the exit face of the crystal, displaying

an on-axis lattice soliton at high nonlinearity (800 V). **c**, Interferogram, showing constructive interference of peak and lobes between the soliton and a plane wave. **d**, Reduction of probe intensity by a factor of eight, at the same voltage as **b** and **c**, results in recovery of discrete diffraction pattern.

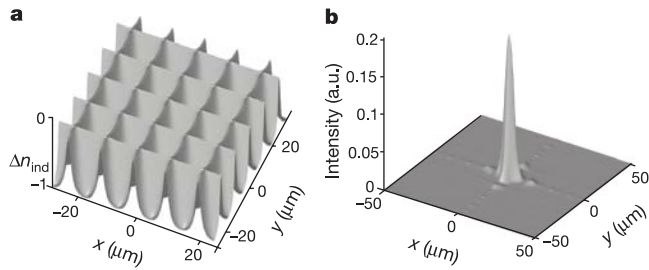


Figure 4 Numerical simulation results depicting the induced photonic lattice and the propagation of a self-trapped staggered wave packet. **a**, Calculated 2D structure of the induced index change (photonic lattice) for the self-defocusing nonlinearity. **b**, Simulated intensity structure of the self-trapped staggered wave packet after 3 cm propagation in our system. As shown here, the energy leakage along the backbone grid is minimal.

is lowered, at the same voltage, the soliton disappears: the beam broadens and the discrete diffraction pattern reappears (Fig. 3d). Because the voltage remains constant, these results imply that the lowered intensity forces the probe beam off the soliton existence curve, that is, the formation of a lattice soliton is an inherently nonlinear effect, which critically depends on the intensity of the soliton beam.

Next, we examine off-axis propagation for which the 'Bloch momentum' of the probe beam lies in the vicinity of the edge of the first Brillouin zone, so that waves propagating in neighbouring waveguides are π out-of-phase with one another. To observe the behaviour, we launch the probe (soliton-forming) beam into a single waveguide of the same array at an angle of 0.55° with respect to the lattice plane. Diffraction in this (angular) region is anomalous (negative)¹³. Hence, self-trapping of a 'bright' wave packet in this regime necessitates a self-defocusing nonlinearity^{11,15}, that is, E_0 has to be negative. The calculated 2D structure of the induced index change (photonic lattice) is shown in Fig. 4a. Notice that, in contrast to the self-focusing case (of Fig. 2), here adjacent waveguides are not fully isolated from one another but are connected through a grid of narrow equipotential 'backbones', which can allow power to leak slowly along this grid. However, this potential can still support

self-trapping of staggered wave packets for a considerable distance, because the energy leakage is very slow. We simulate the propagation of such a self-trapped wave packet, and find it remains a self-trapped entity for very many tunnelling lengths. A typical numerical result is shown in Fig. 4b, which displays the intensity of a self-trapped staggered wave packet after 3 cm propagation in our system. Lines of radiation can be seen along the axes of the array, but their magnitude emphasizes that the power leakage is indeed very small. Our experimental results are shown in Fig. 5. Shown is the intensity of the probe beam as it exits the photonic lattice after propagating 6 mm along the 2D waveguide array. At low voltages, a diffuse diffraction pattern occurs (Fig. 5a), while a staggered (π out-of-phase) self-trapped wave packet is observed (Fig. 5b) at a higher nonlinearity. Indeed, an interferogram (Fig. 5c) confirms this staggered interpretation: the central peak is lowered while the surrounding lobes increase their intensity, indicating destructive and constructive interference, respectively. When the intensity of the probe is lowered, while keeping the voltage across the crystal constant, the diffuse diffraction pattern is recovered (Fig. 5d).

We emphasize that under the same experimental conditions, without the presence of the periodic lattice, no soliton formation is observed. This effect is very pronounced in the staggered case (Fig. 5), where removing the periodic lattice while keeping the negative voltage (self-defocusing nonlinearity) leads to increased expansion of the probe beam, much beyond its natural (linear) diffraction. In the on-axis case (Fig. 3), removing the periodic lattice while keeping the positive voltage shows some small self-focusing of the probe beam, but the ratio between the probe and background beam intensities is off the soliton existence curve and cannot support a soliton^{26–28}. Hence, the soliton formation results presented in Figs 3 and 5, under their experimental conditions, correspond solely to lattice solitons.

The photonic lattice is optically induced by interfering pairs of plane waves in a photosensitive (photorefractive) material, giving our system the exciting capability of fully controlling and reversing both diffraction and nonlinearity. This can be achieved only in dynamical systems such as photorefractive crystals or possibly nematic liquid crystals (for optical waves) or in Bose–Einstein condensates (for matter waves). The technique of optical induction allows for reconfigurable arrays of almost any geometry or Bravais symmetry, including defect states and various three-dimensional

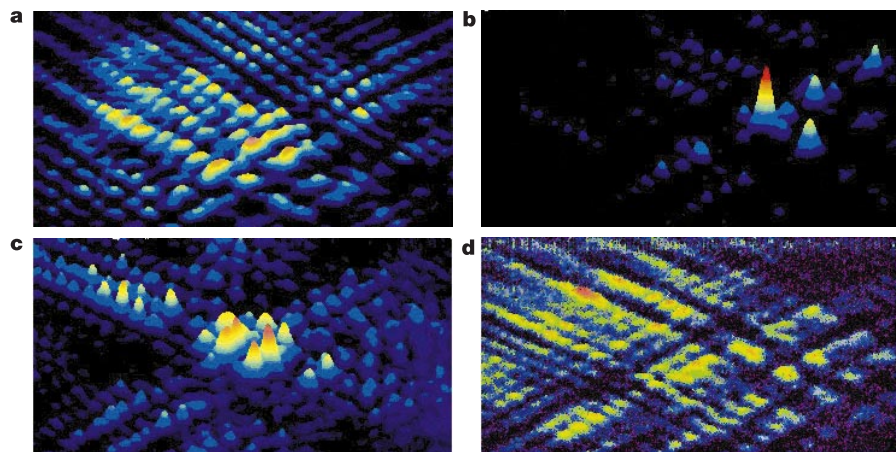


Figure 5 Experimental results presenting the propagation of a probe beam launched into a single waveguide. The angle of the beam is 0.55° with respect to the lattice plane, corresponding to the edge of the first Brillouin zone. **a**, Intensity structure of the probe beam at the exit face of the crystal, displaying diffuse diffraction at low nonlinearity (-200 V). **b**, Intensity structure of the probe beam at the exit face of the crystal,

displaying a self-trapped staggered wave packet at high nonlinearity (-800 V).

c, Interferogram, showing destructive interference of peak and constructive interference of lobes, between the self-trapped wave packet and a plane wave. **d**, Reduction of probe intensity by a factor of eight, at the same voltage as **b** and **c**, results in recovery of diffuse diffraction pattern.

structures. Our results pave the way for the realization of a variety of nonlinear localization phenomena in photonic lattices and crystals^{19–23}. Given the potential of photonic structures for miniaturized optical devices²⁹, we anticipate that lattice solitons, being highly nonlinear entities controlled by light alone³⁰, will become of increasing importance. Finally, our observation directly relates to the proposed lattice solitons in Bose–Einstein condensates⁴, which can be observed in optically induced periodic potentials^{24,25}. □

Received 26 September 2002; accepted 21 January 2003; doi:10.1038/nature01452.

- Davydov, A. S. & Kislukha, N. I. Solitary excitations in one-dimensional molecular chains. *Phys. Status Solidi B* **59**, 465–470 (1973).
- Christodoulides, D. N. & Joseph, R. I. Discrete self-focusing in nonlinear arrays of coupled waveguides. *Opt. Lett.* **13**, 794–796 (1988).
- Su, W. P., Schieffer, J. R. & Heeger, A. J. Solitons in polyacetylene. *Phys. Rev. Lett.* **42**, 1968–1971 (1979).
- Trombettoni, A. & Smerzi, A. Discrete solitons and breathers with dilute Bose–Einstein condensates. *Phys. Rev. Lett.* **86**, 2353–2356 (2001).
- Xie, A., van der Meer, L., Hoff, W. & Austin, R. H. Long-lived amide I vibrational modes in myoglobin. *Phys. Rev. Lett.* **84**, 5435–5438 (2000).
- Trias, E., Mazo, J. J. & Orlando, T. P. Discrete breathers in nonlinear lattices: experimental detection in Josephson junctions. *Phys. Rev. Lett.* **84**, 741–744 (2000).
- Schwartz, U. T., English, L. Q. & Sievers, A. J. Experimental generation and observation of intrinsic localized spin wave modes in an antiferromagnet. *Phys. Rev. Lett.* **83**, 223–226 (1999).
- Swanson, B. I. *et al.* Observation of intrinsically localized modes in a discrete low-dimensional material. *Phys. Rev. Lett.* **82**, 3288–3301 (1999).
- Eisenberg, H. S., Silberberg, Y., Morandotti, R., Boyd, A. R. & Aitchison, J. S. Discrete spatial optical solitons in waveguide arrays. *Phys. Rev. Lett.* **81**, 3383–3386 (1998).
- Morandotti, R., Eisenberg, H. S., Silberberg, Y., Sorel, M. & Aitchison, J. S. Self-focusing and defocusing in waveguide arrays. *Phys. Rev. Lett.* **86**, 3296–3299 (2000).
- Fleischer, J. W., Carmon, T., Segev, M., Efremidis, N. K. & Christodoulides, D. N. Observation of discrete solitons in optically-induced real-time waveguide arrays. *Phys. Rev. Lett.* **90**, 023902 (2003).
- Efremidis, N. K., Sears, S., Christodoulides, D. N., Fleischer, J. W. & Segev, M. Discrete solitons in photorefractive optically-induced photonic lattices. *Phys. Rev. E* **66**, 046602 (2002).
- Eisenberg, H., Silberberg, Y., Morandotti, R. & Aitchison, J. Diffraction management. *Phys. Rev. Lett.* **85**, 1863–1866 (2000).
- Ablovitz, M. J. & Musslimani, Z. H. Discrete diffraction-managed spatial solitons. *Phys. Rev. Lett.* **87**, 254102 (2001).
- Kivshar, Y. S. Self-localization in arrays of defocusing waveguides. *Opt. Lett.* **18**, 1147–1149 (1993).
- Scott, A. C. & Macneil, L. Binding energy versus nonlinearity for a “small” stationary soliton. *Phys. Lett. A* **98**, 87–88 (1983).
- Lederer, F., Darmanyan, S. & Kobayakov, A. in *Spatial Solitons* (eds Trillo, S. & Torruellas, W.) 269–292 (Springer, New York, 2001).
- Malomed, B. A. & Kevrekidis, P. G. Discrete vortex solitons. *Phys. Rev. E* **64**, 026601 (2001).
- Chen, W. & Mills, D. L. Gap solitons and the nonlinear optical response of superlattices. *Phys. Rev. Lett.* **58**, 160–163 (1987).
- Christodoulides, D. N. & Joseph, R. I. Slow Bragg solitons in nonlinear periodic structures. *Phys. Rev. Lett.* **62**, 1746–1749 (1989).
- John, S. & Akozbek, N. Nonlinear optical solitary waves in a photonic band gap. *Phys. Rev. Lett.* **71**, 1168–1171 (1993).
- Mingaleev, S. F. & Kivshar, Y. S. Self-trapping and stable localized modes in nonlinear photonic crystals. *Phys. Rev. Lett.* **86**, 5474–5477 (2001).
- Christodoulides, D. N. & Efremidis, N. K. Discrete temporal solitons along a chain of nonlinear coupled microcavities embedded in photonic crystals. *Opt. Lett.* **27**, 568–570 (2002).
- Anderson, B. P. & Kasevich, M. A. Macroscopic quantum interference from atomic tunnel arrays. *Science* **282**, 1686–1689 (1998).
- Greiner, M., Mandel, O., Hänsch, T. W. & Bloch, I. Collapse and revival of the matter wave field of a Bose–Einstein condensate. *Nature* **419**, 51–54 (2002).
- Segev, M., Crosignani, B., DiPorto, P., Valley, G. C. & Yariv, A. Steady state spatial screening-solitons in photorefractive media with external applied field. *Phys. Rev. Lett.* **73**, 3211–3214 (1994).
- Christodoulides, D. N. & Carvalho, M. I. Bright, dark, and gray spatial soliton states in photorefractive media. *J. Opt. Soc. Am. B* **12**, 1628–1633 (1995).
- Shih, M., Segev, M. & Salamo, G. Circular waveguides induced by two-dimensional bright steady-state photorefractive spatial screening solitons. *Opt. Lett.* **21**, 931–933 (1995).
- Joannopoulos, J. D., Meade, R. D. & Winn, J. N. *Photonic Crystals: Molding the Flow of Light* (Princeton Univ. Press, Princeton, New Jersey, 1995).
- Christodoulides, D. N. & Eugenev, E. D. Blocking and routing discrete solitons in two-dimensional networks of nonlinear waveguide arrays. *Phys. Rev. Lett.* **87**, 233901 (2001).

Acknowledgements This work is part of the MURI programme on optical solitons, and was also supported by the Israeli Science Foundation, and by the German–Israeli DIP project. J.W.F. thanks the Lady Davis Foundation at the Technion for support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.S. (e-mail: msegev@tx.technion.ac.il).

Self-organization of dissolved organic matter to micelle-like microparticles in river water

Martin Kerner^{*†}, Heinz Hohenberg[‡], Siegmund Ertl[§], Marcus Reckermann^{||} & Alejandro Spitz[§]

^{*} University of Hamburg, Institute for Hydrobiology and Fishery Science, D-22765 Hamburg, Zeiseweg 9, Germany

[‡] Heinrich-Pette-Institute for Experimental Virology and Immunology, University of Hamburg, Martinistr. 52, D-20251 Hamburg, Germany

[§] University of Hamburg, Institute for Biogeochemistry and Marine Chemistry, D-20146, Hamburg, Bundesstrasse 55, Germany

^{||} Research and Technology Centre Westcoast, University of Kiel, D-25761 Büsum, Hafentörn, Germany

In aquatic systems, the concept of the ‘microbial loop’ is invoked to describe the conversion of dissolved organic matter to particulate organic matter by bacteria¹. This process mediates the transfer of energy and matter from dissolved organic matter to higher trophic levels, and therefore controls (together with primary production) the productivity of aquatic systems. Here we report experiments on laboratory incubations of sterile filtered river water in which we find that up to 25% of the dissolved organic carbon (DOC) aggregates abiotically to particles of diameter 0.4–0.8 micrometres, at rates similar to bacterial growth. Diffusion drives aggregation of low- to high-molecular-mass DOC and further to larger micelle-like microparticles. The chemical composition of these microparticles suggests their potential use as food by planktonic bacterivores. This pathway is apparent from differences in the stable carbon isotope compositions of picoplankton and the microparticles. A large fraction of dissolved organic matter might therefore be channelled through microparticles directly to higher trophic levels—bypassing the microbial loop—suggesting that current concepts of carbon conversion in aquatic systems require revision.

We studied the capacity for abiotic aggregation of dissolved organic matter (DOM) less than 0.2 µm in diameter present in sterile filtered Elbe water, sampled from the surface at the study site at kilometre 627 (53.543335° N, 9.914018° E) during summer and autumn (June and November 2000) and preserved with 0.02% (w/v) NaN₃, in 5-litre batch experiments under mild agitation in the dark. These experimental procedures ensured diffusion (brownian motion) providing the necessary energy² for aggregation of exclusively natural DOM in the absence of bacteria³ whose numbers remained undetectable below 100 cells ml^{−1} throughout the experiments by transmission electron microscopy (TEM)⁴. Heterotrophic processes were effectively suppressed, because total organic carbon remained constant. Over the experimental period of ten days, abiotic microparticle formation at 24 °C was found to consume about 25% and 7% of the total initial dissolved organic carbon present at similar concentrations in summer and autumn, respectively (Fig. 1a, b). Production rates and abundances of microparticles reached those of bacteria in the river Elbe and in many other rivers and estuaries⁵ (Table 1). Significant amounts of DOC in these environments might therefore be transformed into microparticulate matter by aggregation and not by heterotrophic processes, as is currently assumed.

Flow cytometry revealed that the diameters of the microparticles remain within a narrow range between 0.4 and 0.8 µm. Formation within this restricted size range contrasts the assembly of polymer gels of up to 5 µm described for ocean water under non-turbulent

[†] Present address: SSC Strategic Science Consult Ltd, D-22765 Hamburg, Bodenstedtstr. 16, Germany.

conditions⁶. This difference can be explained by differences in the type of binding. Cation bridging, which was shown to stabilize the marine polysaccharide networks, is not significant for microparticle formation in our experiments, where production decreased by only about 20% in parallel incubations after the addition of 10 mM EDTA as the chelating agent for divalent cations⁶. Non-ionic bonding was further indicated by a 30-fold depletion of Ca^{2+} in the microparticles compared to divalent trace metals (Fig. 2a). Consistent with these findings, TEM images of samples obtained during the incubations showed that fibrils indicative of polymerization of predominantly saccharidic components and capable of forming large networks were only a minor component of the organic microstructures (Fig. 3a–c and Supplementary Information). The microparticles described here thus appear to be completely different from transparent exopolymer particles⁷.

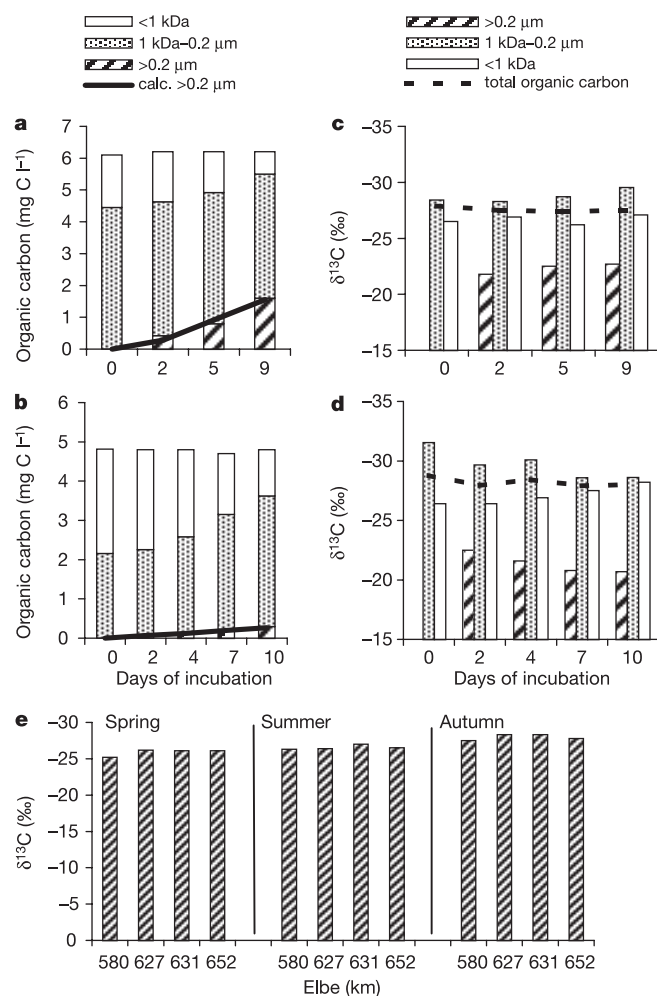


Figure 1 Changes in organic matter size speciation and associated $\delta^{13}\text{C}$ values during abiotic incubation of sterile filtered and poisoned Elbe water. Low-molecular-mass DOC aggregates to high-molecular-mass DOC and further to microparticles ($>0.2\ \mu\text{m}$) both in June (**a, c**) and November (**b, d**) maintaining a constant amount of total organic carbon throughout the incubation. Rate constants for microparticle production (**c**) of $0.7\ \text{day}^{-1}$ (**a**) and $0.056\ \text{day}^{-1}$ (**b**) were calculated with correlation coefficients of $R^2 > 0.96$, assuming a two-step ($A \rightarrow B \rightarrow C$) and one-step ($B \rightarrow C$) first-order reaction (straight lines) in June and November, respectively. **c, d**, Transfer of organic matter between the different molecular weight fractions was followed by changes in $\delta^{13}\text{C}$ contents maintaining constant isotopic balance (dashed line). **e**, For comparison with laboratory results, $\delta^{13}\text{C}$ contents of picoplankton ($0.2\text{--}1.2\ \mu\text{m}$) along the river Elbe between kilometres 580 and 652 at different seasons. (Kilometre 580, for example, is 580 km downstream from the river's source.)

Further insight into the chemical structure of the microparticles was obtained by proton NMR. Results of high-resolution magic-angle-spinning nuclear magnetic resonance (HR-MAS-NMR) (Fig. 2b) revealed a prevalence of mobile CH_2 and CH_3 groups abundant in fatty acids, whereas the sugars and proteins detected in the microparticles (Table 1) did not produce the resonances between 3 and 4 p.p.m. known from NMR studies on humic substances⁸. This finding can only be explained by a significant decrease of their rotational motion; we assume that the matter inside the microparticles is strongly cross-linked, which would have reduced the resonance intensity beyond recognition by extensive line broadening⁹. Long-chain lipids present seemed not to be inhibited in their rotational motion and produced the specific resonances. From these findings we hypothesized a micelle-like structure that was indeed detected for most of the microparticles ($\sim 90\%$) by TEM (Fig. 3b and Supplementary Information). Aggregation as observed for DOC from the Elbe thus shows an extraordinary capacity for self-organization, reaching a level of conformation resembling that of organelles (that is, liposomes) in living cells. Because such well-ordered conformations are indicative of directed processes, it cannot be ruled out that microparticles catalyse very specific biochemical processes associated with compartmentation, selective transport, and information structuring that are known for biological systems¹⁰.

The capacity for aggregation did not depend on the size of the source material (Fig. 1a, b). Aggregation of subcolloidal DOC ($<1\ \text{kDa}$) to colloidal DOC ($>1\ \text{kDa}$) dominated in autumn. In summer, when subcolloidal concentrations were much lower, their aggregation to colloids was exceeded by colloidal aggregation to microparticles. The isotopic data give additional information on the type of aggregating DOC: isotopically heavy subunits of the aggregating pools (that is, subcolloidal DOC in autumn and colloidal DOC in summer) seem to be preferentially involved in the process. In winter, when subcolloidal aggregation was dominant, the colloidal DOC became continuously enriched in ^{13}C at the expense of the subcolloidal ^{13}C . In summer, when colloidal aggregation dominated, the remaining colloidal DOC became depleted in ^{13}C . Microparticles at the final stage of aggregation always reached

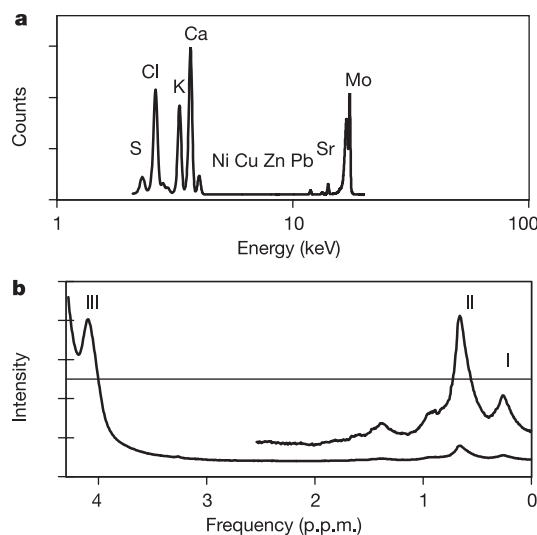


Figure 2 Total reflection X-ray fluorescence difference and proton NMR spectra of microparticles produced during abiotic incubation. **a**, All elements present at similar amounts both in solution and in the microparticles show counts near zero while Ca^{2+} depleted in the microparticles by about 30-fold exhibits a marked X-ray peak. **b**, NMR of microparticles shows major resonances from CH_3 groups (I), CH_2 groups (II) and probably encapsulated water (III) while the diversity of the organic matter detected by destructive analytical methods remained undetected.

Table 1 Characterization of microparticles formed from Elbe River DOM

Parameter	Description	Methods of analysis
Type of reaction	Diffusion controlled aggregation of DOM	Tangential cross-flow filtration; and UV-oxidation (carbon)
Type of binding	Mainly non-ionic	Incubations with and without addition of EDTA; and X-ray-fluorescence
Conformation	Micelle-like	HR-MAS-NMR; and TEM
Size (μm)	0.2 to 2.5	Flow-fluorocytometry after staining with PG, NR, and succinylated Con-A; and TEM
Formation rate at 24 °C (day^{-1})	0.7 (June)/0.06 (November)	UV-oxidation (carbon)
Temperature dependency of formation rate (Q_{10})	5.7 (June)/1.4 (November)	Kinetics of carbon transfer was calculated from parallel laboratory incubations at 11, 18 and 24 °C
Maximum numbers (particles $\times 10^6 \text{ ml}^{-1}$)	32 (June)/1.0 (November)	Flow-fluorocytometry after staining with PG, NR, and succinylated Con-A
Chemical composition	ds DNA, lipids, α -mannose/ α -glucose, acid polysaccharides, proteins	Flow-fluorocytometry after staining with PG, NR, and succinylated Con-A; and colorimetry using BCA (proteins)
Mobile end groups	CH_2 , CH_3	HR-MAS-NMR
$\delta^{13}\text{C}$ -content (‰)	−21 to −23	Sealed tube combustion and mass spectrometry ($\delta^{13}\text{C}$)
C:N ratio	4.1 to 6.8	UV-oxidation (carbon); and wet oxidation (nitrogen)
C-content (fg C per particle)	25 to 55	UV-oxidation (carbon); and flow-fluorocytometry after staining with PG, NR, and succinylated Con-A

$\delta^{13}\text{C}$ values between −21‰ and −23‰ and were thus significantly different from the DOC from which they originated (typically $\delta^{13}\text{C}$ around −28‰). In summary, our results strongly indicate that microparticles include subcolloidal DOC that aggregates into larger and larger clusters.

To study the impact of fresh DOC on the capacity of refractory DOC to aggregate, we incubated in parallel sterile filtered, poisoned Elbe water obtained in autumn and the same water spiked with phytoplankton exudates. These were produced from natural phytoplankton populations of the Elbe dominated by small centric diatoms at logarithmic growth in a flow-through bioreactor¹¹ during addition of isotopically labelled CO_2 ($\delta^{13}\text{C} = -22\text{‰}$) as the substrate. Consistent with the well-known metabolic fractionation by phytoplankton, the exudates had a $\delta^{13}\text{C}$ of −42‰. Addition of these exudates (DOC increase was 10%) increased both the reaction rates and microparticle production by about 40% (data not shown). The fact that the labelled exudates did not modify

the $\delta^{13}\text{C}$ of the microparticles revealed, however, a negligible contribution to microparticle mass. Hence, the function of exudates was limited to that of a trigger of abiotic aggregation of low-molecular-mass DOC, rather than providing the building material, as would be the case in the formation of transparent exopolymer particles^{7,12}.

We found chemical compounds typical of bacteria in the microparticles formed abiotically from Elbe DOM (Table 1). They contained double-stranded DNA, lipids and sugars. Microparticles in natural waters might therefore be confused with bacteria when using the respective fluorescent dyes for detection, which may explain the common finding that the numbers of active bacteria detected in aquatic systems are much lower than the total counts determined by fluorescent techniques¹³. In addition, C:N ratios of between 4.1 and 6.8 and C contents of between 25–55 fg C per particle in the microparticles were similar to values in bacteria¹⁴. This similarity in composition, size and conformation qualifies

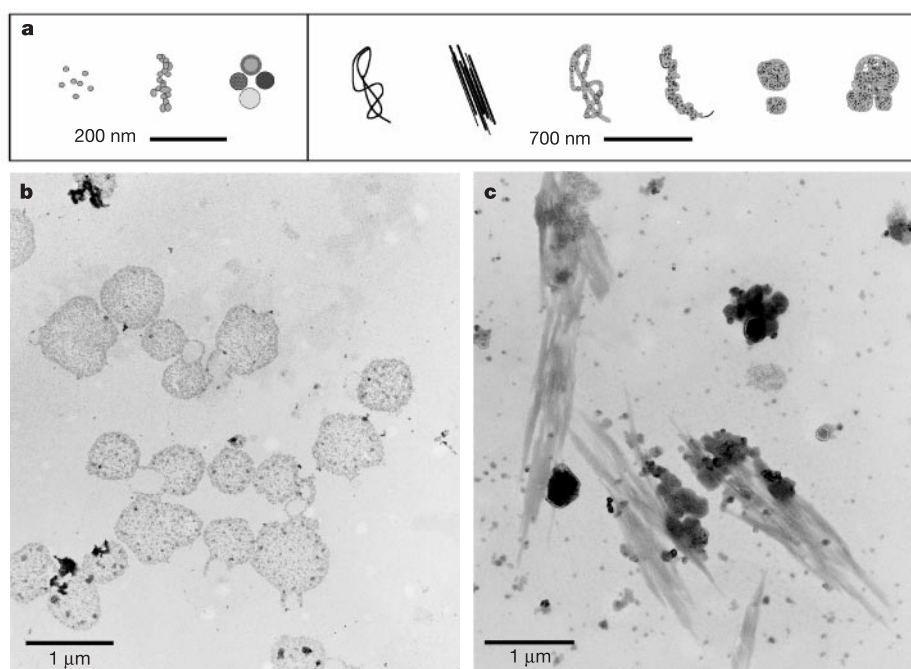


Figure 3 Transmission electron microscopy of aggregated DOM. **a**, Sketched basic structures of all high-molecular-mass organic matter observed during abiotic aggregation of DOM. **b**, TEM image of microparticles and **c**, bundle of fibrillar structures.

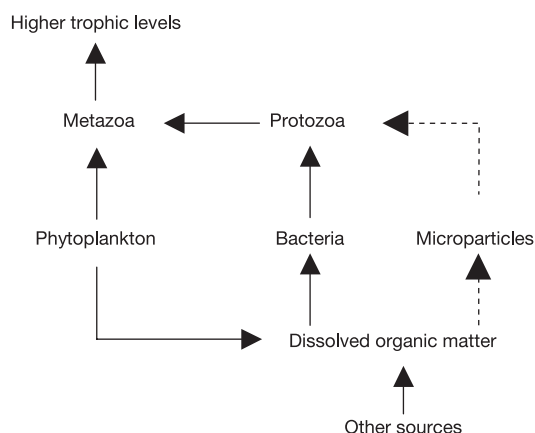


Figure 4 Diagram of DOM transformation in the microbial loop bypassed (dashed lines) by abiotic microparticle formation. The suggested bypass opens access to DOM to bacterivorous organisms whose diet would otherwise be restricted to bacterial biomass.

microparticles as a potentially attractive food source to bacterivorous plankton organisms¹⁵. They might allow growth at rates similar to those of species consuming living matter as a food source. If this is the case, the microbial food web¹ must be extended by the pathway for direct transfer of DOM via abiotic microparticle formation to higher trophic levels (Fig. 4). Our field data (Fig. 1e) strongly suggest this bypass. The absence of the microparticles' distinct isotopic signature (Fig. 1c, d) in the $\delta^{13}\text{C}$ values of the picoplankton (Fig. 1e) means that there must be an efficient microparticle removal process in the field. The $\delta^{13}\text{C}$ values of the picoplankton were similar to those of the total DOC, indicating that the picoplankton consisted mainly of heterotrophic bacteria feeding on DOC because bacteria do not isotopically discriminate against their substrate¹⁶. Bacterial consumption is therefore probably not the removal process, consistent with reports on inhibition of bacterial degradation of artificial liposomes and organic matter adsorbed to beads¹⁷. Another possible removal process is the rapid transformation of microparticles to greater particles. This has not been observed in our laboratory incubations, consistent with the fact that the contact probability between particles is at a minimum in the specific size range of the microparticles². Thus, a rapid removal in the field by further aggregation is also not likely. A removal process fully consistent with our data on $\delta^{13}\text{C}$, microparticle formation rates, size and composition is bacterivorous consumption.

Abiotic transfer of DOC to microparticles would be of great ecological importance because it occurs without carbon losses. It therefore makes organic matter more efficiently available to phagotrophic organisms than heterotrophic bacteria, which have to oxidize about half of the organic carbon to CO_2 to produce energy. Furthermore, abiotic transformation depends only on temperature and the capacity of the DOM to aggregate and is independent of its biological availability. Compounds are transformed which otherwise might resist microbial degradation. Abiotic aggregation therefore has the potential greatly to affect carbon conversion in aquatic systems. This needs to be examined in more detailed studies with the focus on the fate of microparticles. □

Methods

Sampling and incubation

About 50 litres of water collected from the Elbe water surface was sieved first through a $0.125\text{ }\mu\text{m}$ stainless steel mesh and then through a $15\text{ }\mu\text{m}$ nylon mesh to remove larger organisms. This was followed by tangential flow ultrafiltration (Pall Gelman, Minisette) at a pressure of 1 bar, first through $1.2\text{ }\mu\text{m}$ and then through $0.16\text{ }\mu\text{m}$ polyethersulphone membranes (Minisette, Omega). The sterile filtered water was preserved with 0.02% NaN_3 and subsequently pressure-filtered through a $0.2\text{ }\mu\text{m}$ polycarbonate membrane

(Nuclepore). 5 litres of water leaving the filtration cell (Amicon) was directed without contact with the air (to avoid any contamination) into each of three sterile 10-litre glass bottles used for incubations at 11, 18 and 24°C . Control experiments with HgCl_2 (0.01 w/v) as an alternative poison confirmed that NaN_3 was an efficient biocide at the concentration used.

Enrichment of sterile filtered, non-preserved Elbe water with phytoplankton exudates was performed in two parallel flow-through glass bioreactors each with a volume of 2 litres containing natural phytoplankton $<125\text{ }\mu\text{m}$ in diameter from the study site at kilometre 580 (ref. 11). Microscopic examination of the inoculum revealed a phytoplankton composition typical of the Elbe at the study site. The dominant species were small centric diatoms together with *Nitzschia acicularis*, *Cryptomonas* sp., *Asterionella formosa* and *Chlamydomonas* sp. The incubation temperature was 15°C , and the bioreactors were continuously illuminated by light at wavelengths between 295 and 780 nm (Osram lamps, Type 72 and 77) at an intensity of about $500\text{ }\mu\text{mol m}^{-2}\text{ s}^{-1}$, which is typical for the surface of the Elbe on a bright winter day. The medium for phytoplankton incubation was continuously exchanged from the bioreactor and contained 5% CO_2 (with a $\delta^{13}\text{C} = -22\text{‰}$), 20% O_2 , and 75% N_2 . The culture medium was thus richer in CO_2 than the atmosphere, ensuring a constant pH during the incubation of about 7.

Organic carbon

Continuous-flow oxidation of C_{org} to CO_2 was performed using an ultraviolet thin-film-reactor¹⁸. The resulting CO_2 was quantified by a non-dispersive infrared detector (Licor LI-6252). Calibration curves were made with potassium hydrogen phthalate in a range between 0.5 and 8.0 mg Cl^{-1} . The detection limit of the system is 0.1 mg Cl^{-1} , and the precision is $<2\%$ ($n = 5$). POC was quantified by the difference in organic carbon before and after filtration with acid-washed and precombusted Al_2O_3 filters (Anopore, $0.2\text{ }\mu\text{m}$ size exclusion). Separation of DOC $<1\text{ kDa}$ was performed by tangential-flow ultrafiltration.

Particulate organic nitrogen

Wet oxidation to nitrate, that is, peroxydisulphate digestion took place under highly alkaline conditions at a pH > 12 in a sealed glass ampoule at 120°C for 1 h (ref. 19). The detection limit was $30\text{ }\mu\text{g N l}^{-1}$.

Proteins

Microassay method using bicinchoninic acid²⁰.

Flow-cytometry

Size, number, and composition of the particles produced were determined by a BD FACS Vantage flow cytometer equipped with an Innova Enterprise II 621 laser, light regulated at 20 mW ultraviolet, emitting at 488 nm . Particles were characterized by their specific light scatter (forward scatter, FSC, or side scatter, SSC) and green and red fluorescence intensities (green FL1 $530 \pm 20\text{ nm}$; red FL3 $645 \pm 20\text{ nm}$). The particles were counted by determining the fluid volume gravimetrically, which was usually 50 to $100\text{ }\mu\text{l}$. Stains used were: Picogreen (PG, Molecular Probes) which specifically stains double-stranded DNA with a minimum length of 20 base pairs; fluorescein (FITC)-labelled lectin succinyl-concanavalin A (Con-A, Sigma, L 9385) to specifically detect the sugars α -mannose and α -glucose, and benzophenoxazone dye, Nile Red (NR), used as a fluorescent lipid stain²¹. The background fluorescence of all stains used was determined in distilled water and their specific response was tested with negative controls consisting of particles of known composition.

$^{13}\text{C}/^{12}\text{C}$

Sealed-tube combustion of dissolved and particulate organic carbon to CO_2 and subsequent isotope ratio mass spectrometry on a Finnigan MAT 252 (S.E. and A.S., unpublished work). Results were obtained with a precision (preparation and determination) of $\leq 0.5\text{‰}$ at an undetectable procedural blank. Values are given in the δ -notation as deviation (in ‰) to the standard Vienna-PeeDeeBelemnite.

Proton NMR

Magic-angle-spinning measurements with a Bruker Avance DMX 400 MHz spectrometer equipped with a temperature control unit (0.5 K accuracy) using high precession ceramic rotors of 7 mm diameter rotated at $5,000\text{ Hz}$ (ref. 22).

X-ray element analysis

The total reflection X-ray fluorescence difference spectrum was obtained by the subtraction of the spectra for the elements present in Elbe water and on microparticles produced therein, both normalized by their Zn^{2+} content as an internal standard²³.

Transmission electron microscopy

TEM analyses were performed on samples obtained at the beginning, during, and at the end of the experiments, prepared by a directional drying technique (DDT): drying of an unstained subsample of $0.6\text{ }\mu\text{l}$ with a soft, warm, humid and directional air flow on a very thin carbon support film of $\sim 3\text{ nm}$ thickness made hydrophilic by glow discharge (details in Supplementary Information). At the beginning the samples were free of visible structures. During and at the end of the incubations similar fine organic structures and microparticles were identified. To ensure that all organic structures present were imaged after preparation by DDT we applied a whole-mount drying technique (WMDT) on parallel samples using a negative/positive metal stain: mixing of $1\text{ }\mu\text{l}$ of a subsample with 0.002% w/v uranyl acetate and drying of the mixture as a whole onto the same hydrophilic carbon film used for DDT. Using WMDT we imaged the same structures found by DDT, but about 30% of the grid was covered with dark spots resulting from densely stained

layers of superimposed material. Staining was not necessary to image even fibrils of <3 nm in thickness which until now were thought to become visible by TEM only by advanced staining techniques using thicker supporting films than we used²⁴ (see Supplementary Information). Thin films are known to increase the specific contrast of the matter and allow us to image very fine structures²⁵. In addition, experiments with varying concentrations of the metal stain revealed that we could image microparticles by WMDT and DDT only by applying very low or no concentrations of metal stain, respectively. They thus seem to be very sensitive to staining salts. DDT is superior to WMDT in that it allowed a controlled attachment and separation of the organic matter of different densities and thus microparticles were shown to exist separately from the remaining organic structures. Furthermore, DDT avoids problematic preparative steps like ultrafiltration, diafiltration, freeze-drying and rehydration.

Received 4 November 2002; accepted 27 January 2003; doi:10.1038/nature01469.

1. Azam, F. *et al.* The ecological role of water column microbes in the sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983).
2. Kepkay, P. E. Particle aggregation and the biological reactivity of colloids. *Mar. Ecol. Prog. Ser.* **109**, 293–304 (1994).
3. Riley, G. A. Organic aggregates in seawater and the dynamics of their formation and utilization. *Limnol. Oceanogr.* **4**, 372–381 (1963).
4. Hohenberg, H., Mannweiler, K. & Müller, M. High-pressure freezing of cell suspensions in cellulose capillary tubes. *J. Microsc.* **175**, 34–43 (1994).
5. Ducklow, H. W. & Shiah, F.-K. In *Aquatic Microbiology* (ed. Ford, T. E.) 261–287 (Blackwell, Oxford, 1993).
6. Chin, W. C., Orellana, M. V. & Verdugo, P. Spontaneous assembly of marine dissolved organic matter into polymer gels. *Nature* **391**, 568–572 (1998).
7. Leppard, G. G. The characterization of algal and microbial mucilages and their aggregates in aquatic ecosystems. *Sci. Tot. Environ.* **165**, 103–131 (1995).
8. Hertkorn, N. *et al.* Comparative analysis of partial structures of a peat humic and fulvic acid using one and two dimensional nuclear magnetic resonance spectroscopy. *J. Environ. Qual.* **31**, 375–387 (2002).
9. Nanny, M. A. & Maza, J. P. Noncovalent interactions between monoaromatic compounds and dissolved humic acids: a deuterium NMR T1 relaxation study. *Environ. Sci. Technol.* **35**, 379–384 (2001).
10. Pollard, T. D. & Earnshaw, W. C. *Cell Biology* 1st edn 277–281 (Elsevier Science, Philadelphia, 2002).
11. Kerner, M. & Spitz, A. Nitrate regeneration coupled to degradation of different size fractions of DON by the picoplankton in the Elbe estuary. *Microb. Ecol.* **41**, 69–81 (2001).
12. Passow, U. & Alldredge, A. L. A dye binding assay for the spectrophotometric measurement of transparent exopolymer particles (TEP). *Limnol. Oceanogr.* **40**, 1326–1335 (1995).
13. Bianchi, A. & Guiliano, L. Enumeration of viable bacteria in the marine pelagic environment. *Appl. Environ. Microbiol.* **62**, 174–177 (1996).
14. Fukuda, R., Ogawa, H., Nagata, T. & Koike, J. Direct determination of carbon and nitrogen contents of natural bacterial assemblages in marine environments. *Appl. Environ. Microbiol.* **65**, 3352–3358 (1998).
15. Posch, T. & Arndt, H. Uptake of sub-micrometre- and micrometre-sized detrital particles by bacterivorous and omnivorous ciliates. *Aquat. Microb. Ecol.* **10**, 45–53 (1996).
16. Hullar, M., Fry, B., Peterson, B. J. & Wright, R. T. Microbial utilization of estuarine dissolved organic carbon: a stable isotope tracer approach tested by mass balance. *Appl. Environ. Microbiol.* **62**, 2489–2493 (1996).
17. Nagata, T. & Kirchman, D. L. Roles of submicron particles and colloids in microbial food webs and biogeochemical cycles within marine environments. *Adv. Microb. Ecol.* **15**, 81–103 (1997).
18. Huber, S. A. & Frimmel, F. H. Flow injection analysis of organic and inorganic carbon in the low-ppb range. *Anal. Chem.* **63**, 2122–2130 (1991).
19. Raimbault, P. & Slawyk, G. A semi-automatic, wet-oxidation method for the determination of particulate organic nitrogen collected on filters. *Limnol. Oceanogr.* **36**, 405–408 (1991).
20. Smith, P. K. *et al.* Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150**, 76–85 (1985).
21. Greenspan, P. & Fowler, S. D. Spectrofluorometric studies of the lipid probe, Nile red. *J. Lip. Res.* **26**, 781–789 (1985).
22. Volke, F. *et al.* Characterisation of antibiotic moenomycin A interaction with phospholipid model membranes. *Chem. Phys. Lipids* **85**, 115–123 (1997).
23. Prange, A., Tuempling, W. Jr, Niedergesaess, R. & Jantzen, E. The whole river Elbe in one view: Element distribution of the river Elbe from the spring to the mouth. *Wasserw. Wassertechn.* **7**, 22–31 (1995).
24. Santschi, P. H. *et al.* Fibrillar polysaccharides in marine macromolecular organic matter, as imaged by atomic force microscopy and transmission electron microscopy. *Limnol. Oceanogr.* **43**, 896–908 (1998).
25. Harris, J. R. *Negative Staining and Cryoelectron Microscopy: The Thin Film Technique* 12–21 (BIOS Scientific Publishers, Oxford, UK, 1997).

Supplementary Information accompanies the paper on Nature's website
(<http://www.nature.com/nature>).

Acknowledgements We thank L. Ehrhardt for technical assistance and H. Lother for discussions. This research was funded by the Deutsche Forschungsgemeinschaft (M.K. and A.S.). The Heinrich-Pette-Institute is supported by the Bundesministerium für Gesundheit und die Freie und Hansestadt Hamburg.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K.
(e-mail: kerner@uni-hamburg.de).

Core formation in planetesimals triggered by permeable flow

Takashi Yoshino, Michael J. Walter & Tomoo Katsura

Institute for Study of the Earth's Interior, Okayama University, Yamada 827, Misasa, Tottori-ken 682-0193, Japan

The tungsten isotope composition of meteorites indicates that core formation in planetesimals occurred within a few million years of Solar System formation^{1,2}. But core formation requires a mechanism for segregating metal, and the 'wetting' properties of molten iron alloy in an olivine-rich matrix is thought to preclude segregation by permeable flow unless the silicate itself is partially molten^{3–5}. Excess liquid metal over a percolation threshold, however, can potentially create permeability in a solid matrix, thereby permitting segregation. Here we report the percolation threshold for molten iron–sulphur compounds of approximately 5 vol.% in solid olivine, based on electrical conductivity measurements made *in situ* at high pressure and temperature. We conclude that heating within planetesimals by decay of short-lived radionuclides can increase temperature sufficiently above the iron–sulphur melting point ($\sim 1,000^\circ\text{C}$) to trigger segregation of iron alloy by permeable flow within the short time-frame indicated by tungsten isotopes. We infer that planetesimals with radii greater than about 30 km and larger planetary embryos are expected to have formed cores very early, and these objects would have contained much of the mass in the terrestrial region of the protoplanetary nebula. The Earth and other terrestrial planets are likely therefore to have formed by accretion of previously differentiated planetesimals, and Earth's core may accordingly be viewed as a blended composite of pre-formed cores.

Static percolation theory predicts that permeability forms at any melt fraction when the dihedral angle—the interfacial angle formed at solid–melt triple junctions—is less than 60° , whereas above 60° melt can migrate only if the melt fraction reaches a percolation threshold⁶. Dihedral angles measured for molten Fe alloy in a solid olivine-rich matrix are high, around 90 – 100° (refs 3–5), effectively rendering the silicate impermeable. Thus, core formation in planetesimals may have required considerable amounts of silicate melting to 'lubricate' the system, enhancing the mobility of Fe alloy⁷. Melting of ultramafic silicate rocks requires heating to temperatures in excess of $\sim 1,200^\circ\text{C}$, whereas the eutectic melting temperature in the Fe–FeS system is lower, around $1,000^\circ\text{C}$ (refs 8–10). It is necessary to determine the value of the percolation threshold for S-bearing Fe alloys to evaluate the physical mechanism and temperature required for core segregation.

The percolation threshold in solid–liquid systems with large dihedral angles ($\sim 100^\circ$) has not been accurately determined. Numerical models show a low percolation threshold of 5.7 vol.% at a dihedral angle of 100° (ref. 6) if melt is homogeneously distributed in the solid matrix. But if melt-free grain edges are energetically favoured, melt coalesces into a heterogeneous distribution of isolated pockets^{7,11} and the percolation threshold could be as high as 40 vol.%. The percolation threshold cannot be accurately estimated only on the basis of measured dihedral angles, so direct experimental determination is required.

We performed *in situ* electrical conductivity measurements at high pressure and temperature on mixtures of olivine and molten Fe–S compounds to determine the percolation threshold. The connectivity of small amounts of alloy melt can be detected with this technique, because the electrical conductivity of Fe–S compounds (10^5 – 10^8 S m^{-1}) is nearly ten orders of magnitude

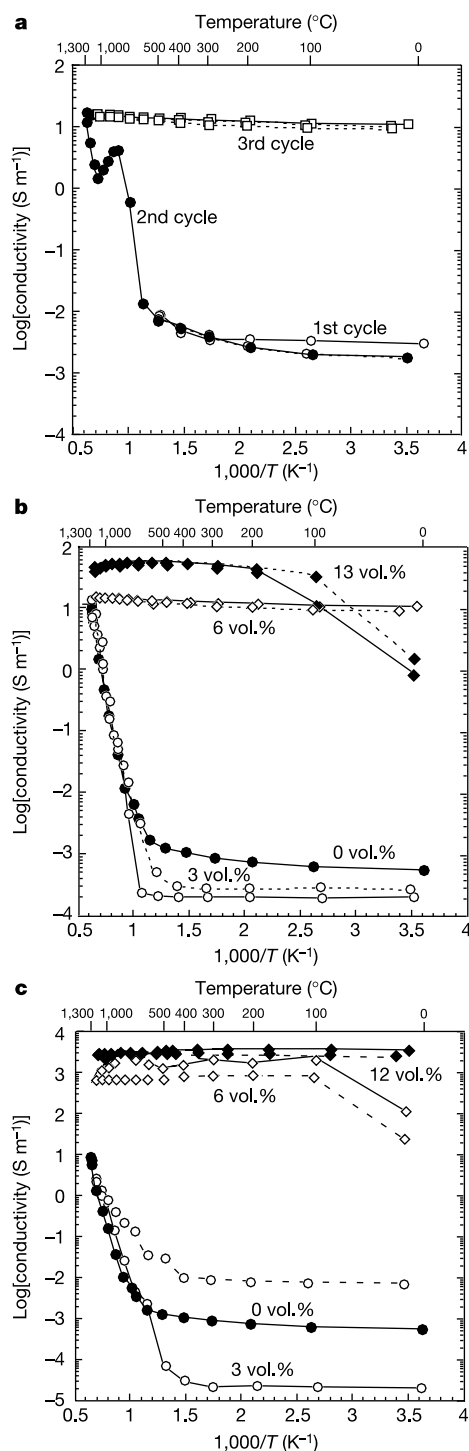


Figure 1 Electrical conductivity of olivine-sulphide mixtures as a function of reciprocal temperature. Lines and dashed lines indicate heating and cooling cycles, respectively. **a**, Electrical conductivity of the olivine-FeS mixture with a ratio of 94:6 in volume. Open circles, filled circles and open squares correspond to conductivity values measured in the first, second and third heating-cooling cycles, respectively. Note that the trace of the third heating-cooling cycle is consistent with that of the second cooling cycle, which is obscured in the plot by the points from the third cycle. **b**, Electrical conductivity of FeS-bearing starting material during the third heating-cooling cycle. The filled squares, open circles, open diamonds and filled diamonds correspond to mixtures containing 0, 3, 6 and 13 vol.% FeS, respectively. **c**, Electrical conductivity of Fe₆₄S₃₆-bearing starting material during the third heating-cooling cycle. The filled squares, open circles, open diamonds and filled diamonds correspond to mixtures containing 0, 3, 6 and 12 vol.% Fe₆₄S₃₆, respectively. The steep linear trends in the high-temperature region of samples with 0 and 3 vol.% metal are attributed to the BN sample container.

greater than for silicates ($<10^{-3} \text{ S m}^{-1}$), even at high temperature. Starting materials were composed of San Carlos olivine mixed in variable proportions with alloy of two compositions, FeS and Fe₆₄S₃₆ (see Methods), the latter corresponding approximately to the eutectic composition (Fe₆₁S₃₉) in Fe-FeS at 3 GPa (ref. 9). High pressures and temperatures were generated using a cubic high-pressure apparatus. Sample resistance was measured by means of a pseudo-four-wire method, the principle of which has been described elsewhere¹².

Experiments were conducted at 3 GPa, and three heating-cooling cycles were performed in each run. During the first cycle, the sample was heated to 500 °C, annealed for 30 min, then cooled to room temperature. During the second cycle, the sample was heated to the maximum temperature (1,200 °C for Fe₆₄S₃₆ or 1,300 °C for FeS), which was above the eutectic melting point in Fe-FeS ($<1,000 \text{ °C}$)⁹, but below the melting point of Fe₉₀ olivine ($\sim 1,800 \text{ °C}$). The sample was held at maximum temperature for over 20 h to achieve textural equilibrium, and then cooled to room temperature. During the third cycle, the sample was again heated to the maximum temperature, annealed for 30 min, then cooled to room temperature. During each cycle temperature was changed in 100 °C increments, and electrical conductivity was measured at each step.

An example of conductivity measurements along heating and cooling paths is shown in Fig. 1a (run no. 44). In the first cycle, conductivity was low and almost constant (10^{-2} S m^{-1}). Textural equilibrium was not achieved. During the heating portion of the second cycle, conductivity increased by more than three orders of magnitude from 600 °C (10^{-2} S m^{-1}) to 1,300 °C (10 S m^{-1}), and remained nearly constant at constant temperature. Molten Fe alloy achieves a textural steady state in the silicate matrix during this cycle, which remains during subsequent heating and cooling (see Supplementary Information). Sample conductivity in the third cycle was repeatable without significant hysteresis.

Experimental results are shown in Fig. 1b and c for FeS and Fe₆₄S₃₆, respectively. The characteristics of the temperature-conductivity paths are similar for the two Fe-alloy compositions, although absolute conductivities are different. In runs with Fe-alloy proportions above 6 vol.%, the conductivity was high and nearly independent of temperature during the third cycle. Preservation of high conductivity after temperature was reduced below the eutectic melting point indicates that the Fe alloy was connected when molten. In runs with lower metal proportions (~ 3 and 0 vol.%), conductivities were low and nearly constant up to 600 °C, but increased considerably as temperature was raised to the maximum temperature. Temperature-conductivity paths in these runs are essentially the same as in a reference run made without Fe-alloy melt (see Methods), in which low conductivity was re-established upon cooling during all cycles, and we conclude that Fe-alloy melt was not connected. Our results bracket the percolation threshold between 3 and 6 vol. % independent of the sulphur content of the

Table 1 Summary of experiments

Run no.	S (mol%) [*]	wt% [†]	ϕ [‡]	Duration (h) [§]	σ_{maxT} (S m^{-1})	σ_{500} (S m^{-1}) [¶]
AMA040	50	20	0.13	20	55.62	54.58
AMA044	50	10	0.06	72	17.89	12.21
AMA048	50	5	0.03	40	1.11	0.0003
AMA049	-	0	0	6	1.07	0.0008
AMA071	36	20	0.12	66.4	2,738	2,896
AMA069	36	10	0.06	65	817	835
AMA064	36	5	0.03	48	0.63	0.0003

^{*} Mole per cent sulphur in Fe-S compound added to starting material. During experiments the actual sulphur content was slightly lowered relative to these values owing to dissociation of a trace Mo component in the sulphide phase.

[†] Weight per cent Fe-S compound in each sample.

[‡] Calculated volume fraction of the Fe-S phase based on density of olivine (3.3 g cm^{-3}), FeS (5.5 g cm^{-3}) and Fe₆₄S₃₆ (6.3 g cm^{-3}).

[§] Duration time (h) at the maximum temperature (1,300 °C for AMA040-049; 1,200 °C for AMA064-069) during the second heating-cooling cycle.

^{||} Electrical conductivity at 1,200 °C measured at the onset of the second cooling cycle.

[¶] Electrical conductivity at 500 °C measured during the second cooling cycle.

iron alloy (Table 1). This is much lower than reported previously (10–25 vol.%) for an olivine-metal system based on electrical conductivity measurements made at ambient conditions¹³.

The percolation threshold may be sensitive to the composition of the Fe alloy. For example, it has been inferred from undifferentiated meteorites that Fe–S–O melts in an olivine matrix under high oxygen and sulphur fugacities have lower dihedral angles (60–90°)^{5,14} than do reduced iron alloys (100–110°)³. Although we have explored only a limited compositional range, our measurements indicate no dependence of percolation threshold on sulphur content of the Fe alloy. Dihedral angles in our runs were large ($95 \pm 5^\circ$), and the percolation threshold is consistent with predictions from numerical models for pinch-off boundaries⁶.

Connectivity implies the potential for permeable flow, although metal segregation was not observed in our runs owing to surface tension effects. Whether melt forms two-dimensional sheets along grain boundaries or one-dimensional tubes along triple junctions is important for a quantitative understanding of permeability. Two-dimensional cross-sections of the quenched run products show a predominance of melt pools at grain corners and tubes along triple junctions, indicating that melt connectivity is established by one-dimensional tubes.

Our results show that Fe–S alloy would segregate at temperatures high enough to melt the alloy ($\sim 1,000^\circ\text{C}$) but below the silicate solidus, reducing considerably the amount of heating needed for core formation. The migration rate of metal due to gravitational acceleration is proportional to the planetary radius. For a set of realistic parameters for grain size (10^{-3} m), metal fraction (10%), density difference between metal and silicate ($3,500\text{ kg m}^{-3}$), gravitational acceleration (0.1 m s^{-2}), viscosity (0.005 Pa s), and porosity exponent of permeability ($n = 1-3$), the migration velocity at the surface of a planetesimal with a radius of 100 km is 10^0 to 10^2 myr^{-1} , based on Darcy's law. Thus, if permeable flow is established, segregation velocities are rapid in comparison with the timescale of core formation predicted from the Hf–W isotope system^{1,2}.

The volume fraction of metal in primitive chondritic meteorites is highly variable, from 0.1 to 20 vol.% (ref. 15), and the amount of iron alloy as a function of heliocentric distance in the protoplanetary nebula during planetesimal accretion is not known accurately. Accretion theory indicates that Earth is an amalgamation of smaller planetesimals, and so may be useful as a model for average metal content in objects originating within its 'feeding zone' ($\sim 0.5-2.5$ astronomical units from the Sun, where 1 AU is the average Sun–Earth distance)¹⁶. Earth's metallic core comprises about 15 vol.% of Earth, and we use this value as a working estimate to model metal segregation in planetesimals. Thus, for a percolation threshold of $\sim 5\text{ vol.}\%$, about two-thirds of the available metal would segregate to form a core as soon as Fe alloy melts.

The source of heat necessary to raise the temperature of a planetesimal sufficiently to melt either Fe alloy or silicate is not known. The decay of short-lived radioactive nuclei (^{26}Al and ^{60}Fe) active early in the Solar System is a primary candidate¹⁷⁻¹⁹. Other possibilities include inductive heat gained from early solar outflow and heat from impacting bodies²⁰. Planetary cooling via heat conduction is very efficient for small objects, but the centre of an object with a radius greater than about 30 km retains most of its heat. Accretion models show that in an initial stage of dust coagulation a large number of planetesimals with radii from 1 to 10 km form in the terrestrial region of the protoplanetary nebula²¹. In a second stage, effects of gravity, energy equipartition and gas drag drive a period of runaway growth yielding an array of objects that include planetesimals and planetary embryos with radii spanning up to thousands of kilometres (ref. 22). These events occur within 500,000 yr or so of nebula formation.

Here, we estimate the thermal evolution in an accreted planetesimal with CI chondritic chemical composition caused by decay of ^{26}Al and ^{60}Fe , as this heat source may have been uniformly present in accreting material. We use a range of initial ratios for $^{26}\text{Al}/^{27}\text{Al}$ and $^{60}\text{Fe}/^{56}\text{Fe}$ of 5.6×10^{-6} to 9.8×10^{-6} and 2×10^{-6} to 2×10^{-8} , respectively¹⁷⁻¹⁹. Our calculation is illustrated in Fig. 2. The W isotope composition of differentiated meteorites

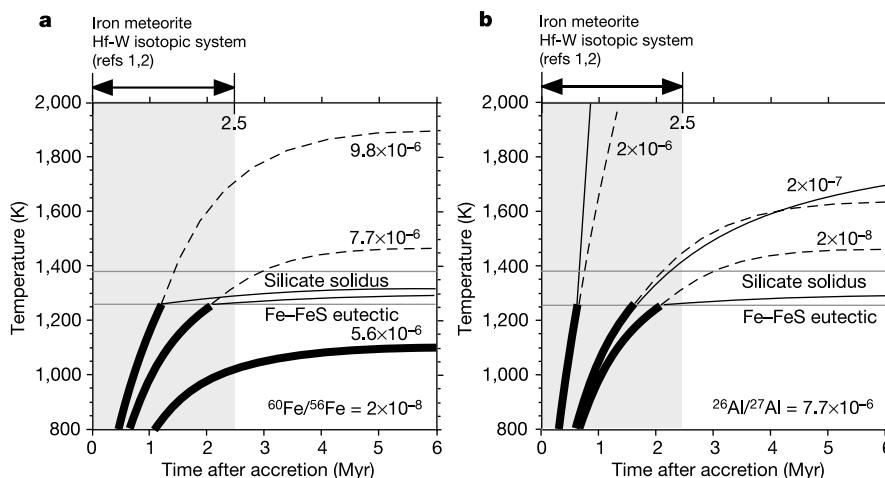


Figure 2 A thermal model for post-accretion heating via decay of ^{26}Al and ^{60}Fe in a planetesimal of CI chondritic chemical composition and with a radius of over 30 km. The silicate:metal proportion is 85:15 (vol.%). An initial temperature of 255 K corresponds to the present radiation equilibrium temperature of Earth. The solidi of Fe–FeS (ref. 8) and Allende silicate (ref. 10) at 1 atm are shown. The W isotope composition of iron meteorites places a limit on the timing of core formation of about 3 Myr after Solar System formation. Although not strictly constrained, we assume an accretion timescale of $5 \times 10^5\text{ yr}$ (ref. 22). We use a range of initial ratios for $^{26}\text{Al}/^{27}\text{Al}$ and $^{60}\text{Fe}/^{56}\text{Fe}$ of 5.6×10^{-6} to 9.8×10^{-6} and 2×10^{-6} to 2×10^{-8} , respectively¹⁷⁻¹⁹. Thick lines represent heating paths for the silicate + metal mixture. Thin solid and dashed lines show the heating paths after core segregation for the metal (eutectic melt composition) and silicate portions,

respectively. **a**, Effect on planetesimal heating for variable $^{26}\text{Al}/^{27}\text{Al}$ (numbers on curves) at a low value of $^{60}\text{Fe}/^{56}\text{Fe}$. For the lowest value of $^{26}\text{Al}/^{27}\text{Al}$, the Fe–FeS solidus is not reached. For values of $^{26}\text{Al}/^{27}\text{Al}$ higher than $\sim 6.6 \times 10^{-6}$, the Fe–FeS solidus is reached and core formation can occur by permeable flow within 2.5 Myr. For the highest value of $^{26}\text{Al}/^{27}\text{Al}$, the silicate solidus is also reached within this timeframe and any remaining metal would be efficiently scavenged to the core. **b**, Effect on planetesimal heating for variable $^{60}\text{Fe}/^{56}\text{Fe}$ (numbers on curves) at an intermediate value of $^{26}\text{Al}/^{27}\text{Al}$. All heating paths reach the Fe–FeS solidus within 2.5 Myr. For values of $^{60}\text{Fe}/^{56}\text{Fe}$ greater than $\sim 10^{-7}$ continued radioactive heating after metal segregation causes temperatures in the core to reach well above the silicate solidus. This considerable core superheat could cause melting of the silicate in planetesimals, resulting in further differentiation and volcanism.

indicate that core formation occurred in planetesimals within about 3 Myr (refs 1, 2) of Solar System formation. If the lowest initial ratios are used for both radionuclides, the temperature would not reach the Fe-FeS melting point and core segregation could not occur in the absence of another heat source (Fig. 2a). However, in general the heat provided by radioactive decay is sufficient to heat a planetesimal to the Fe-FeS melting point, resulting in metal segregation within the timeframe allowed by W isotopes. If high initial ratios are used for both radionuclides then the silicate melting point could be reached as well, and any trapped Fe alloy remaining after initial segregation would also segregate. Another consequence of a high initial $^{60}\text{Fe}/^{56}\text{Fe}$ is considerable heating in the core of the planet after metal segregation. The amount of superheat produced can be prodigious, and is a plausible heat source for early volcanism on planetesimals and internal differentiation as indicated by ordinary chondrite meteorites²⁰.

Our results indicate that planetesimals with radii greater than about 30 km and larger planetary embryos are expected to have formed cores very early, and these objects would have contained much of the mass in the terrestrial region of the protoplanetary nebula²². The Earth and other terrestrial planets are likely to have formed by accretion of previously differentiated planetesimals, and Earth's core may be viewed as a blended composite of pre-formed cores. □

Methods

Materials and experimental techniques

San Carlos olivine was mixed with reagent-grade FeS in volume ratios of 100:0, 97:3, 94:6 and 87:13, and with $\text{Fe}_{64}\text{S}_{36}$ in volume ratios of 100:0, 97:3, 94:6 and 88:12. The mixtures were finely ground and most grains were 5–10 μm in diameter, with occasional larger grains (10–100 μm). The sample was packed into a boron nitride (BN) capsule, and was 1 mm in diameter and 3 mm long. The BN capsule was surrounded by molybdenum foil so that the oxygen fugacity would be close to the Mo-MoO₂ buffer, and was placed at the centre of a graphite heating element. Graphite electrodes were placed in contact with the sample, and measured resistance included the electrodes in addition to the sample. A reference resistance (10^2 or $10^3 \Omega$) was connected to the sample in series, and a sinusoidal signal with an amplitude of 1 V and a frequency in the range 0.1–0.01 Hz was applied to the circuit. Voltages on the sample and reference were measured simultaneously. The sample resistance was calculated from that of the reference and the ratio of voltage on the sample to that on the reference. Sample conductivity was calculated from the resistance and the sample dimensions. We measured the electrical conductivity of polycrystalline olivine without Fe-S compounds to establish a reference for the experimental configuration. The conductivity of the reference was very low ($\sim 10^{-3} \text{ S m}^{-1}$) up to nearly 700 °C, and then rapidly increased with temperature to 1,300 °C (9 S m^{-1}). The electrical conductivity from 700 to 1,300 °C is not attributed to olivine, which has a much lower conductivity at 1,300 °C ($3 \times 10^{-3} \text{ S m}^{-1}$)²³, but to the BN capsule. The temperature-conductivity path for this reference established a baseline below which values are not derived from the sample.

Received 10 September 2002; accepted 28 January 2003; doi:10.1038/nature01459.

1. Yin, Q. *et al.* A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* **418**, 949–952 (2002).
2. Kleine, J., Munker, C., Mezger, K. & Palme, H. Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf-W chronometry. *Nature* **418**, 952–955 (2002).
3. Shannon, M. C. & Agee, C. B. High pressure constraints on percolative core formation. *Geophys. Res. Lett.* **23**, 2717–2720 (1996).
4. Minarik, W. G., Ryerson, F. J. & Watson, E. B. Textural entrapment of core-forming melts. *Science* **272**, 530–533 (1996).
5. Ballhaus, C. & Ellis, D. J. Mobility of core melts during Earth's accretion. *Earth Planet. Sci. Lett.* **143**, 137–145 (1996).
6. von Bagen, N. & Waff, H. S. Permeabilities, interfacial areas and curvatures of partially molten systems: Results of numerical computations of equilibrium microstructures. *J. Geophys. Res.* **91**, 9261–9276 (1986).
7. Taylor, G. J. Core formation in asteroids. *J. Geophys. Res.* **97**, 231–249 (1992).
8. Fei, Y., Bertka, C. M. & Finger, L. High-pressure iron-sulfur compound, Fe_3S_2 , and melting relations in the Fe-FeS system. *Science* **275**, 1621–1623 (1997).
9. Usselman, T. M. Experimental approach to the state of the core. Part I: The liquidus relations of the Fe-rich portion of the Fe-Ni-S system from 30 to 100 kb. *Am. J. Sci.* **275**, 278–290 (1975).
10. Agee, C. B. Melting temperatures of the Allende meteorite: Implications for a Hadean magma ocean. *Phys. Earth Planet. Inter.* **100**, 41–47 (1997).
11. Stevenson, D. J. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 231–249 (Oxford Univ. Press, Oxford, 1990).
12. Katsura, T., Sato, K. & Ito, E. Electrically conductivity measurement of minerals at high pressures and high temperatures. *Rev. High Press. Sci. Technol.* **7**, 18–21 (1998).
13. Jurewicz, S. R. & Jones, J. H. Preliminary results of olivine/metal wetting experiments and the direct measurement of metal phase interconnectivity. *Proc. Lunar Planet. Sci. Conf.* **26**, 709–710 (1995).
14. Gaetani, G. A. & Grove, T. L. Wetting of mantle olivine by sulfide melt: implications for Re/Os ratios in mantle peridotite and late-stage core formation. *Earth Planet. Sci. Lett.* **169**, 147–163 (1999).

15. Scott, E. R. D., Love, S. G., Krot, A. N. in *Chondrules and the Protoplanetary Disk* (eds Hewins, R. H. *et al.*) 87–96 (Cambridge Univ. Press, Cambridge, 1996).
16. Wetherill, G. W. Provenance of the terrestrial planets. *Geochim. Cosmochim. Acta* **58**, 4513–4520 (1994).
17. Hutcheon, I. D. & Hutchinson, R. Evidence from the Semarkona ordinary chondrite for ^{26}Al heating of small planets. *Nature* **337**, 238–241 (1989).
18. Shukolyukov, A. & Lugmair, G. W. Live iron-60 in the early solar system. *Science* **259**, 1138–1142 (1993).
19. Shukolyukov, A. & Lugmair, G. W. ^{60}Fe - ^{60}Ni systematics in the eucrite Caldera. *Meteoritics* **31**, A129 (1997).
20. Taylor, S. R. *Solar System Evolution: A New Perspective* (Cambridge Univ. Press, Cambridge, 1992).
21. Weidenschilling, S. J. The origin of comets in the solar nebula: a unified model. *Icarus* **127**, 290–306 (1997).
22. Kokubo, E. & Ida, S. Formation of protoplanets and planetesimals in the solar nebula. *Icarus* **143**, 15–27 (2000).
23. Xu, Y., Poe, B. T., Shankland, T. J. & Rubie, D. C. Electrical conductivity of olivine, wadsleyite and ringwoodite under upper-mantle conditions. *Science* **280**, 1415–1418 (1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank E. Ito, K. T. Koga and T. Hatakeyama for discussions; O. Nishikawa and C. Oka for technical assistance; and D. Kohlstedt and W. Minarik for comments and suggestions that improved the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.Y. (e-mail: tyoshino@misasa.okayama-u.ac.jp).

Escalation of a coevolutionary arms race through host rejection of brood parasitic young

Naomi E. Langmore*, Sarah Hunt† & Rebecca M. Kilner‡

* School of Botany and Zoology, Australian National University, Canberra, Australian Capital Territory 0200, Australia

† School of Biological Sciences, University of Bristol, Woodland Road, Clifton, Bristol, BS8 1UG, UK

‡ Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK

Cuckoo nestlings that evict all other young from the nest soon after hatching impose a high reproductive cost on their hosts¹. In defence, hosts have coevolved strategies to prevent brood parasitism. Puzzlingly, they do not extend beyond the egg stage^{2–5}. Thus, hosts adept at recognizing foreign eggs remain vulnerable to exploitation by cuckoo nestlings^{6,7}. Here we show that the breach of host egg defences by cuckoos creates a new stage in the coevolutionary cycle. We found that defences used during the egg-laying period by host superb fairy-wrens (*Malurus cyaneus*) are easily evaded by the Horsfield's bronze-cuckoo (*Chrysococcyx basalis*), a specialist fairy-wren brood parasite. However, although hosts never deserted their own broods, they later abandoned 40% of nests containing a lone Horsfield's bronze-cuckoo nestling, and 100% of nests with a lone shining bronze-cuckoo nestling (*Chrysococcyx lucidus*), an occasional fairy-wren brood parasite. Our experiments demonstrate that host discrimination against victor-cuckoo nestlings is possible, and suggest that it has selected for the evolution of nestling mimicry in bronze-cuckoos.

The Horsfield's bronze-cuckoo is an Australian brood parasite that lays a white egg covered with fine red-brown speckling, closely resembling the eggs of its fairy-wren (Maluridae) and thornbill (Acanthizidae) hosts (Fig. 1a). Within 48 h of hatching, the cuckoo

nestling evicts host eggs or young⁸ and so becomes the sole occupant of the nest. At our study site in Campbell Park, Canberra, Australia, 19–37% of superb fairy-wren nests per year contained an egg of the Horsfield's bronze-cuckoo (1999–2002, $N = 349$ nests in which eggs were laid). The fairy-wren's response depended on the timing of parasitism. Cuckoo eggs laid before the host commenced laying were usually sewn into the nest lining (13 out of 15), ensuring that they would fail to hatch, whereas those laid during the host-laying period were almost all accepted (52 out of 53; before versus during comparison, G -statistic test, $G^2 = 47.5$, $P < 0.0001$). When the cuckoo added her egg after the host had begun incubation, the clutch was usually deserted (7 out of 8; during versus after comparison $G^2 = 31.5$, $P < 0.0001$), but when we placed cuckoo eggs in the nest after incubation had begun they were accepted (26 out of 27, natural parasitism versus relocated eggs, $G^2 = 23.1$, $P < 0.0001$). Desertion was therefore triggered by the presence of an adult female cuckoo that was more likely to be caught at the nest by hosts during incubation than during egg laying. The presence of the cuckoo egg itself did not provoke desertion. The effective parasitism rate (excluding buried or deserted cuckoo eggs) ranged from 16–32% per year.

In further experiments, we tested whether fairy-wrens ever discriminated against foreign eggs. During the egg-laying period, we added to the nest a painted egg that differed from the fairy-wren's in colour, pattern or size (see Methods). Fairy-wrens did not reject odd eggs on the basis of colour or pattern (3 out of 16 blue, 2 out of 11 spotted, 4 out of 18 fairy-wren colour, 0 out of 7 brown painted real eggs deserted or rejected; $\chi^2 = 1.8$, 3 degrees of freedom (d.f.), $P = 0.62$; see also ref. 8), but they were more likely to desert clutches containing an egg larger than their own (7 out of 15 big versus 1 out

of 11 small parchment-coloured model eggs deserted or rejected; $G^2 = 4.2$, $P = 0.04$). The dim interior of the fairy-wren's domed nest (mean plus standard error = 976 ± 105 lx, $N = 107$ nests) may have promoted tactile, rather than visual recognition of foreign eggs, a strategy seen in other dome-nesting host species^{5,9}. However, it is a poor defence against the Horsfield's bronze-cuckoo, which lays a relatively small egg¹⁰, similar in size to fairy-wren eggs (Fig. 1b). In summary, our data show that fairy-wrens can mount some defences against parasitism around the time of egg laying but they do not reject or desert foreign eggs by their appearance. Providing that the Horsfield's bronze-cuckoo female times her egg laying to coincide with that of her targeted host, she will beat host defences at this stage of the nesting cycle.

Acceptance of cuckoo eggs can result in a stable end to the arms race^{11–13} but if parasitism rates are consistently high, hosts would benefit from the evolution of new defences against cuckoo parasitism. The coevolutionary arms race between superb fairy-wrens and Horsfield's bronze-cuckoos has escalated to the stage of discrimination against cuckoo chicks. Through observations at unmanipulated nests, we found that superb fairy-wrens deserted 11 out of 29 (11 out of 42 including depredated nests) Horsfield's bronze-cuckoo nestlings (see also ref. 14) but they never abandoned broods of their own young ($N = 95$ nests, 18 with daily nest watches). The likelihood of cuckoo chick desertion was not influenced by whether any host young had hatched before eviction ($N = 37$, $G^2 = 0.83$, 1 d.f., $P = 0.36$), nor was there a time-of-season effect (desertion rate against month September to January, $G^2 = 1.9$, 4 d.f., $P = 0.7$). In all 11 cases of desertion, and usually when the cuckoo was 3–6 days old ($N = 7$), the female stopped provisioning the cuckoo and, while it was still alive, began building a new nest and soliciting copulations. Although the male(s) in the group sometimes continued feeding the cuckoo for up to three days longer (5 cases), it eventually died from cold or starvation. The corpse was then typically devoured by meat ants *Iridomyrmex purpureus*. Although previous studies of other brood parasitic species have found that some non-evicting parasitic chicks are less effective at eliciting care from their foster parents than host nestlings^{15–17}, ours shows unambiguously that hosts can reject alien young in their nest.

To test how fairy-wren females recognized a cuckoo in their nest,

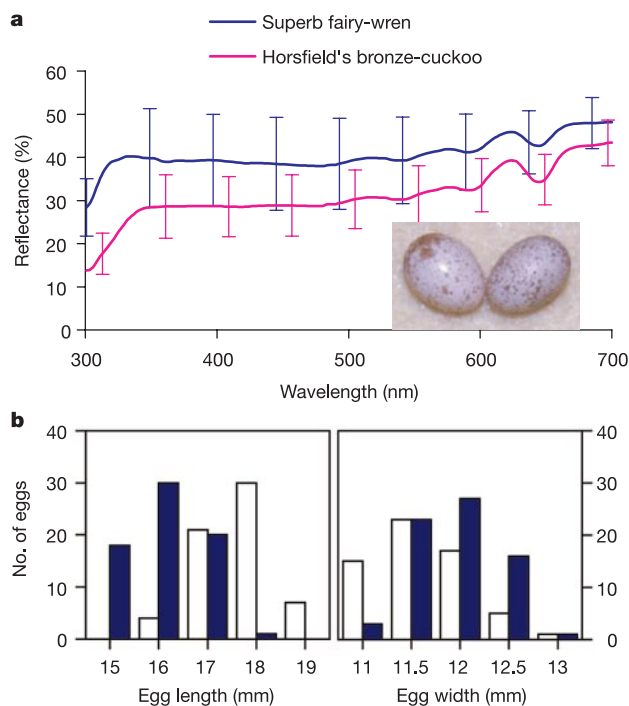


Figure 1 Comparisons of Horsfield's bronze-cuckoo and fairy-wren eggs. **a**, Mean ($\pm$ s.d.) spectral reflectance of Horsfield's bronze-cuckoo eggs (pink line, $N = 9$) and superb fairy-wren eggs (blue line, $N = 26$ from 16 clutches). Standard deviations represent variation within and between eggs. There was a close colour match between cuckoo and host eggs. Cuckoo eggs were slightly less reflective (principal components analysis²⁷, 9 eggs per species, PC1: $F_{1,16} = 7.80$, $P = 0.013$), probably because they have more dark speckling at the narrower end. The photograph shows a superb fairy-wren egg (left) and a Horsfield's bronze-cuckoo egg (right). **b**, Distribution of egg length and width for Horsfield's bronze-cuckoos (white bars) and superb fairy-wrens (blue bars).

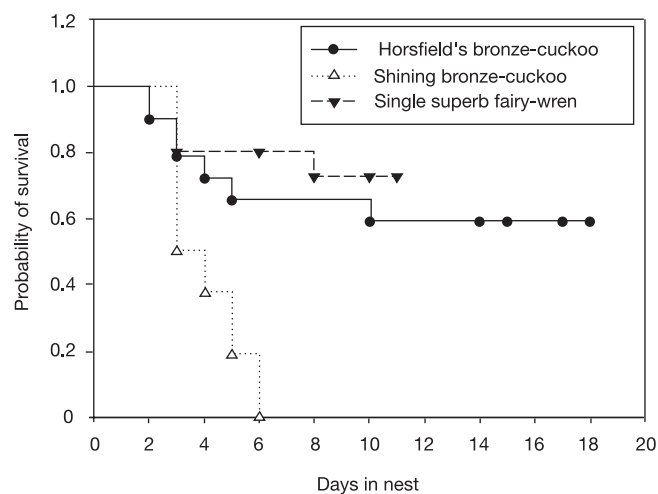


Figure 2 Kaplan-Meier estimate of the survival functions of Horsfield's bronze-cuckoo chicks, shining bronze-cuckoo chicks and lone superb fairy-wren chicks in superb fairy-wren nests. The probability of survival differed significantly between the three species ($\chi^2 = 9.01$, 2 d.f., $P = 0.01$). Comparing probability of survival to that of a single fairy-wren, there was no difference in the risk ratio for Horsfield's bronze-cuckoos (risk ratio = 0.82, 95% confidence limits 0.4, 1.59), whereas there was a significant difference for shining bronze-cuckoos (risk ratio = 2.44, 95% confidence limits 1.21, 4.85).

we manipulated clutches (see Methods) so that after hatching, nests contained either a lone Horsfield's bronze-cuckoo chick ($N = 20$), a lone shining bronze-cuckoo chick ($N = 8$), or a lone fairy-wren chick ($N = 14$). Desertion occurred in all three treatments, which suggests that, in part, females recognize cuckoos simply because they are alone in the nest. However, other cues must be important because shining bronze-cuckoo chicks were always abandoned, whereas single fairy-wren chicks and Horsfield's bronze-cuckoos were deserted at a significantly lower and similar rate (Fig. 2).

We suggest that chick desertion by hosts has evolved in response to cuckoo parasitism. Our experiment shows that cuckoo chick desertion is not simply a by-product of a life history strategy to avoid wasting time on single chick broods. If broods containing a single chick were unprofitable, abandonment of lone chicks should have occurred with equal frequency for all three species. Nor is the consistent desertion of any lone chick a strategy practised by a subset of the female population. Ten females were parasitized by lone chicks of two species; two out of five that reared a lone fairy-wren chick subsequently deserted a Horsfield's bronze-cuckoo chick, and five out of five reared a Horsfield's bronze-cuckoo chick but deserted a shining bronze-cuckoo chick. Furthermore, chick desertion cannot be a pre-existing strategy against non-kin in the nest because host females take the decision to abandon the cuckoo even though they are always related to their entire brood at unparasitized nests ($N = 40$ broods¹⁸), whereas males commonly are not¹⁸.

As the rare fairy-wren parasite (shining bronze-cuckoo) was routinely recognized by hosts, whereas the specialist parasite (Horsfield's bronze-cuckoo) was less frequently abandoned, host defences

against cuckoo chicks appear to have caused reciprocal adaptations in cuckoo nestlings. Preliminary experimental evidence suggests that hosts do not use visual cues to distinguish cuckoo chicks. Shining bronze-cuckoo chicks hatch in one of two distinct colour morphs: a pinkish-yellow morph, which appears similar to fairy-wren chicks (Fig. 3b), or a black morph (Fig. 3c), which differs markedly from host young. Horsfield's bronze-cuckoos chicks are intermediate in colour (Fig. 3a). Despite their similarity to host young, pale morph shining bronze-cuckoo chicks were always deserted by hosts ($N = 4$), whereas the more alien looking Horsfield's bronze-cuckoo chicks were often accepted (14 out of 20; Fisher's exact test, $P = 0.02$).

We suggest that discrimination of vocal cues by hosts may have selected for mimetic begging calls in Horsfield's bronze-cuckoos. We quantified species differences in begging call structure from sonograms of nestling vocalizations, by measuring the frequency of the loudest part of the call, its length and its frequency range, and found that the begging calls of shining bronze-cuckoos mimic those of fairy-wren chicks less well than those of Horsfield's bronze-cuckoo chicks (Fig. 3; see also ref. 19). Neither peak frequency (analysis of variance (ANOVA) species $\times$ age interaction: $F_{2,21} = 0.40$, $P = 0.66$) nor frequency range (ANOVA species $\times$ age interaction: $F_{2,21} = 0.40$, $P = 0.66$) differed significantly between the three species as chicks grew older. However, shining bronze-cuckoo chicks uttered longer calls than fairy-wren offspring, and they became increasingly different from fairy-wren calls as chicks grew older (ANOVA species $\times$ age interaction: $F_{2,21} = 8.20$, $P = 0.002$; shining bronze-cuckoo versus superb fairy-wren, Fisher protected least significant difference (PLSD), $P = 0.03$), whereas Horsfield's bronze-cuckoos did not differ significantly from fairy-wrens in this respect (Fisher PLSD, $P = 0.40$).

We have yet to determine whether a host's ability to discriminate against cuckoo chicks is innate or acquired. Great reed warbler *Acrocephalus arundinaceus* cuckoo hosts imprint on the appearance of their first clutch during egg laying and thereafter reject any eggs that appear odd by comparison⁴. However, similarly learnt nestling recognition cannot evolve among hosts of ejector cuckoos because the cost of misimprinting on a lone cuckoo chick, and ever after rejecting host young, outweighs the benefits of avoiding exploitation⁶. According to Lotem's hypothesis⁶, females that accept cuckoos must have misimprinted on them during their first breeding attempt and should subsequently reject all other young. We found no evidence to support this view. Females that accepted a Horsfield's bronze-cuckoo nestling did not abandon a lone fairy-wren chick in a later breeding attempt (0 desertions in six cases, binomial test, $P = 0.03$).

Why are superb fairy-wrens apparently unique among hosts of ejector cuckoos in their ability to desert cuckoo nestlings? The cost of accepting a parasitic chick is proportional to the likelihood of finding a cuckoo chick in the nest. Therefore it depends on the incidence of parasitism and the strength of host defences at the egg stage, and so may be greater for fairy-wrens than for other hosts. For example, British reed warbler *Acrocephalus scirpaceus* hosts of the European cuckoo *Cuculus canorus* reject roughly 20% of cuckoo eggs added to their nests, which means that 5–13% of hosts end up rearing a cuckoo chick³. By contrast, 13–37% of superb fairy-wren nests monitored at three populations across New South Wales contained Horsfield's bronze-cuckoo eggs (this study; see also refs 20, 21). Even allowing for mistimed egg laying by female Horsfield's bronze-cuckoos, and consequent nest abandonment by hosts, fairy-wrens seem more likely to encounter a cuckoo chick than British reed warblers.

Furthermore, fairy-wren hosts may stand to gain greater net benefits than other hosts by abandoning cuckoo young. The breeding season of superb fairy-wrens is longer (three–six months²²) than that of a typical European cuckoo host (for example, reed warbler, 2.5 months²³), so fairy-wrens are better able to capitalize on chick

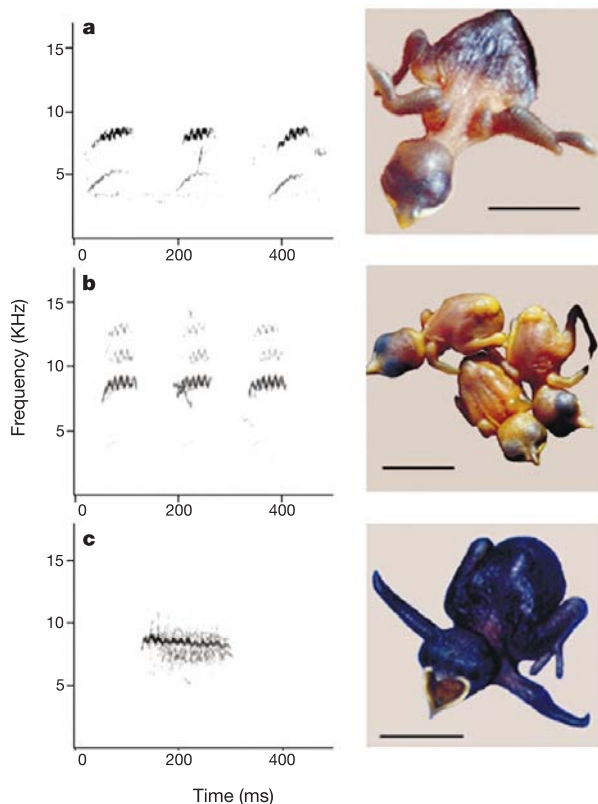


Figure 3 Sonograms from begging calls recorded six days after hatching from a Horsfield's bronze-cuckoo chick (**a**), a brood of three superb fairy-wren chicks (**b**) and a shining bronze-cuckoo chick (**c**). Begging calls of nestling Horsfield's bronze-cuckoos ($N = 7$ recordings from 4 chicks), shining bronze-cuckoos ($N = 5$ recordings from 2 chicks) and superb fairy-wren broods ($N = 14$ recordings from 8 broods) were analysed with Canary 1.2.1. Photographs show the same species 1–2 days after hatching. The photograph in **b** shows two superb fairy-wren chicks (left and right) with a pale morph shining bronze-cuckoo chick (centre). Scale bars, 10 mm.

desertion by re-nesting. In addition, fairy-wrens might more easily bear the costs, revealed by our experiments, of mistakenly rejecting lone host young. For example, the incidence of host broods with a single chick among superb fairy-wren (3.08% of 1,587 broods; A. Cockburn, personal communication) and reed warbler (3.8% of 2,094 broods^{3,24}) broods is similar, and both species rear one to three fledglings per nest^{22,23}. However, fairy-wrens can rear up to three broods per year (34.4% of females rear two broods successfully²⁵), yielding 2.6–3.9 fledglings on average^{14,20,21,25}. By contrast, only 8–32% of reed warblers have a second brood²³. Thus a single chick represents a relatively smaller fraction of reproductive success for a fairy-wren than a reed warbler.

In short, we predict that cuckoo chick discrimination will evolve only when cuckoos have completely evaded host defences at the egg stage, and only then when parasitism rates are sufficiently high to outweigh the costs of recognition errors, breeding seasons are sufficiently long to allow re-nesting, and hosts have sufficiently high fecundity to bear the cost of mistakenly deserting single chicks of their own. □

Methods

Study species

Cooperatively breeding superb fairy-wren groups comprise up to four males and one breeding female, who alone builds the nest and incubates the clutch²². They are a primary host of the Horsfield's bronze-cuckoo²⁶. The shining bronze-cuckoo and the fan-tailed cuckoo (*Cacomantis flabelliformis*), occasional fairy-wren brood parasites²⁶, were both present at our study site but neither parasitized our ringed population during the study.

Egg colour measurement

We measured spectral reflectance of unincubated host and cuckoo eggs using a narrow-ended (2.5-mm diameter) UV-VIS unidirectional reflectance probe, attached to a USB2000 spectroradiometer and PX-2 Xenon strobe (all from Ocean Optics Inc). During measuring, eggs were replaced with model eggs, which were removed when real eggs were returned. We took six measurements at randomly chosen locations within each of three regions (base, middle and tip; 18 readings per egg). The probe illuminated areas approximately 1.5 mm in diameter. Reflectance was calculated relative to a Spectralon 99% reflectance standard between 300 and 700 nm. To prevent egg damage, we added a rubber tip, which extended 2 mm beyond the end of the steel-cased probe. The probe was held perpendicular to the surface of the egg and measurements were taken under a light-proof sheet. Owing to the fineness of the speckle, all reflectance spectra include both white background (spectrally flat) and brown speckle (increasing reflectance across the spectrum, with triple peaks at long wavelengths).

Egg experiments

Both model and real eggs were used in 100 egg discrimination tests ($N = 63$ female fairy-wrens). Model eggs were made of Alumilite super plastic ($N = 22$ small (fairy-wren size: 16×12 mm) blue; 11 small parchment; 15 large (fan-tailed cuckoo size: 20×15 mm) parchment), cast in silicone moulds and painted with Daler-Rowney or Plaid acrylic paints (titanium white, cobalt, turquoise, buckskin brown, nutmeg, coffee bean, parchment, moss green). Real eggs ($N = 18$ blue; 12 brown; 18 parchment base, fairy-wren-like brown speckling; 7 parchment base, large brown spotting) were used within 14 days of laying and obtained from captive zebra finches (*Taeniopygia guttata*, $N = 7$, egg size 15×11 mm) or superb fairy-wren clutches ($N = 43$ eggs) that had been abandoned or where the clutch size had been reduced for the single chick experiment (see below). A single egg was added to the nest during egg laying ($N = 82$) or early incubation ($N = 18$, 0–9 days into incubation, median of 3 days). No female experienced the same treatment twice or was subjected to more than two treatments. Each nest was used only once. We checked the nest the following day and again five days later to determine the outcome. Depredated nests were excluded. There was no significant difference in the response of hosts to plastic compared to real eggs (3 out of 16 blue real eggs compared to 3 out of 22 blue plastic fairy-wren-sized eggs rejected or deserted; $G^2 = 0.2$, $P = 0.67$), nor was the outcome affected by the timing of parasitism (two rejected out of ten blue real eggs placed during laying compared to one out of seven during incubation; $\chi^2 = 0.03$, 1 d.f., $P < 0.99$). We caged 124 out of 734 nests found with garden mesh during incubation. This significantly reduced nest depredation (28% caged, 66% uncaged; $\chi^2 = 46.2$, 1 d.f., $P < 0.0001$) and did not cause chick desertion: we performed daily nest watches on 15 superb fairy-wren broods in caged nests and none was deserted. The light availability inside active fairy-wren nests was measured between 11:00 and 13:00 Eastern Standard Time in full sunlight using a Kyoritsu illuminometer 5200.

Chick experiments

Chick desertion was not caused by interference at the nest. Twenty-six nests containing cuckoo chicks were checked once on hatching day and thereafter their fate was monitored with 1-h nest watches conducted almost daily roughly 50 m from the nest. The nest was not approached again until the chick was deserted ($N = 10$), preyed on ($N = 9$) or fledged ($N = 7$). Eighteen control fairy-wren nests containing unmanipulated broods were subjected to the same treatment and none was deserted.

To create broods of single fairy-wrens, we reduced clutches to a single egg plus two plastic model fairy-wren eggs during incubation. Model eggs were removed within 48 h of hatching. Shining bronze-cuckoo eggs were found in nests of yellow-rumped (*Acanthiza chrysorrhoa*) or buff-rumped thornbills (*A. reguloides*) and transferred to fairy-wren nests before or during incubation. Horsfield's bronze-cuckoo eggs were similarly transferred between fairy-wren nests. After hatching, daily nest watches were essential to detect desertion, because corpses were quickly dispatched by meat ants and thereafter desertion could not be reliably distinguished from depredation. Our experiments were conducted with ethics approval from the Australian National University Animal Experimentation Ethics Committee.

Nestling calls were recorded for 45 min per day, 3–6 days after hatching, with a Sony tie-clip microphone (ECM-T7), clipped within 30 cm of the nest, and connected by a 5-m extension cable to a Sony Professional Walkman (WM-D6C).

Received 18 November 2002; accepted 21 January 2003; doi:10.1038/nature01460.

1. Davies, N. B. *Cuckoos, Cowbirds and Other Cheats* (Poyser, London, 2000).
2. Brooke, M. de L. & Davies, N. B. Egg mimicry by cuckoos *Cuculus canorus* in relation to discrimination by hosts. *Nature* **335**, 630–632 (1988).
3. Davies, N. B. & Brooke, M. de L. Cuckoos versus reed warblers: adaptations and counteradaptations. *Anim. Behav.* **36**, 262–284 (1988).
4. Lotem, A., Nakamura, H. & Zahavi, A. Constraints on egg discrimination and cuckoo host coevolution. *Anim. Behav.* **49**, 1185–1209 (1995).
5. Marchetti, K. Egg rejection in a passerine bird: size does matter. *Anim. Behav.* **59**, 877–883 (2000).
6. Lotem, A. Learning to recognize nestlings is maladaptive for cuckoo *Cuculus canorus* hosts. *Nature* **362**, 743–745 (1993).
7. Kilner, R. M., Noble, D. G. & Davies, N. B. Signals of need in parent-offspring communication and their exploitation by the common cuckoo. *Nature* **397**, 667–672 (1999).
8. Brooker, M. G. & Brooker, L. C. The comparative breeding behaviour of two sympatric cuckoos, Horsfield's bronze-cuckoo *Chrysococcyx basalis* and the shining bronze-cuckoo *C. lucidus*, in Western Australia: a new model for the evolution of egg morphology and host specificity in avian brood parasites. *Ibis* **131**, 528–547 (1989).
9. Mason, P. & Rothstein, S. I. Coevolution and avian brood parasitism: Cowbird eggs show evolutionary response to host discrimination. *Evolution* **40**, 1207–1214 (1986).
10. Krüger, O. & Davies, N. B. The evolution of cuckoo parasitism: a comparative analysis. *Proc. R. Soc. Lond. B* **269**, 375–381 (2002).
11. Davies, N. B. & Brooke, M. de L. An experimental study of co-evolution between the cuckoo, *Cuculus canorus*, and its hosts. II. Host egg markings, chick discrimination and general discussion. *J. Anim. Ecol.* **58**, 225–236 (1989).
12. Davies, N. B., Brooke, M. de L. & Kacelnik, A. Recognition errors and probability of parasitism determine whether reed warblers should accept or reject mimetic cuckoo eggs. *Proc. R. Soc. Lond. B* **263**, 925–931 (1996).
13. Lotem, A. & Nakamura, H. in *Parasitic Birds and Their Hosts. Studies in Coevolution* (eds Rothstein, S. I. & Robinson, S. K.) 223–235 (Oxford Univ. Press, Oxford, 1998).
14. Mulder, R. A. *Evolutionary Ecology of the Mating System of Superb Fairy-wrens*. PhD thesis, Australian National Univ. (1992).
15. Fraga, R. M. in *Parasitic Birds and their Hosts. Studies in Coevolution* (eds Rothstein, S. I. & Robinson, S. K.) 173–193 (Oxford Univ. Press, Oxford, 1998).
16. Lichtenstein, G. Low success of shiny cowbird chicks parasitizing rufous-bellied thrushes: chick-chick competition or parental discrimination? *Anim. Behav.* **61**, 401–413 (2001).
17. Payne, R. B., Woods, J. L. & Payne, L. L. Parental care in estrildid finches: experimental tests of a model of *Vidua* brood parasitism. *Anim. Behav.* **62**, 473–483 (2001).
18. Mulder, R. A., Dunn, P. O., Cockburn, A., Lazenby-Cohen, K. A. & Howell, M. J. Helpers liberate female fairy-wrens from constraints on extra-pair mate choice. *Proc. R. Soc. Lond. B* **255**, 223–229 (1994).
19. Payne, R. B. & Payne, L. L. in *Parasitic Birds and their Hosts. Studies in Coevolution* (eds Rothstein, S. I. & Robinson, S. K.) 152–169 (Oxford Univ. Press, Oxford, 1998).
20. Tidemann, S. C. *The Behavioural Ecology of Three Coexisting Fairy-wrens (Maluridae: Malurus)*. PhD thesis, Australian National Univ. (1983).
21. Nias, R. C. *Co-operative Breeding in the Superb Fairy-wren Malurus cyaneus*. PhD thesis, Univ. New England (1987).
22. Rowley, I. & Russell, E. *Fairy-wrens and Grasswrens* (Oxford Univ. Press, Oxford, 1997).
23. Ogilvie, M. A. in *Handbook of the Birds of Europe, the Middle East and North Africa, The Birds of the Western Palearctic* Vol. VI (eds Cramp, S. & Brooks, D. J.) 193–212 (Oxford Univ. Press, Oxford, 1992).
24. Bibby, C. J. Some breeding statistics of Reed and Sedge Warblers. *Bird Study* **25**, 207–222 (1978).
25. Rowley, I. The life history of the superb blue wren *Malurus cyaneus*. *Emu* **64**, 251–297 (1965).
26. Brooker, M. G. & Brooker, L. C. Cuckoo hosts in Australia. *Aust. Zool. Rev.* **2**, 1–67 (1989).
27. Cuthill, I. C., Bennett, A. T. D., Partridge, J. C. & Meier, E. J. Plumage reflectance and the objective assessment of avian sexual dichromatism. *Am. Nat.* **153**, 183–200 (1999).

Acknowledgements N.E.L. was supported by an Australian Research Council Australian Postdoctoral Fellowship. R.M.K. was supported by a Royal Society Dorothy Hodgkin Research Fellowship and a Royal Society University Research Fellowship. Spectrophotometric equipment was purchased with a BBSRC grant. We thank Environment Australia for allowing us to work at Campbell Park; S. Butchart, A. Cockburn, N. Davies, M. Hall, G. Maurer, J. Madden, N. Macgregor and A. Peters for assistance in the field; A. Cockburn for statistical advice; and M. de la Brooke, A. Cockburn, N. Davies, R. Heinsohn, C. Hinde, A. Lotem, R. Magrath and S. Rothstein for comments on the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.E.L. (e-mail: naomi.langmore@anu.edu.au).

Growth dynamics underlying petal shape and asymmetry

Anne-Gaëlle Rolland-Lagan^{*†}, J. Andrew Bangham^{*} & Enrico Coen[†]

^{*} School of Information Systems, UEA, Norwich, NR4 7TJ, UK

[†] John Innes Centre, Norwich Research Park Colney, Norwich, NR4 7UH, UK

Development commonly involves the generation of complex shapes from simpler ones. One way of following this process is to use landmarks to track the fate of particular points in a developing organ^{1–7}, but this is limited by the time over which it can be monitored. Here we use an alternative method, clonal analysis⁸, whereby dividing cells are genetically marked and their descendants identified visually, to observe the development of *Antirrhinum* (snapdragon) petals. Clonal analysis has previously been used to estimate growth parameters of leaves^{9–11} and *Drosophila* wings^{12–14} but these results were not integrated within a dynamic growth model. Here we develop such a model and use it to show that a key aspect of shape—petal asymmetry—in the petal lobe of *Antirrhinum* depends on the direction of growth rather than regional differences in growth rate. The direction of growth is maintained parallel to the proximodistal axis of the flower, irrespective of changes in shape, implying that long-range signals orient growth along the petal as a whole. Such signals may provide a general mechanism for orienting growth in other growing structures.

The transformations in shape of a growing structure depend on the growth properties of its component regions. For each region and time point, these properties can be conveniently captured by three

types of parameter: the rate of increase in size (growth rate), the degree to which growth occurs preferentially in any direction (anisotropy) and the angle at which the principal direction of growth is oriented relative to an underlying coordinate system (direction) (Fig. 1a). These parameters can vary both in space (that is, between regions) and time, so there are potentially many different ways, with varying degrees of complexity, by which one shape can be transformed into another. Describing the growth of any structure therefore requires parameters to be determined experimentally. Because growth is a complex spatiotemporal problem, the quantitative contribution of these parameters can only be evaluated effectively by incorporating them within a dynamic growth model. This can establish whether the parameters are sufficient to account for the observed shape changes and allows the contribution of each parameter to be explored.

One problem when using clonal analysis to measure growth parameters is that the orientation of growth cannot be inferred simply from the final clone shape¹⁵ because the internal coordinate system of the organ will often be deformed by growth. Relating a mature clone to its growth orientation at the time of initiation can only be done when this deformation is known. We have compared clones induced at successive times during development to determine growth parameters together with the sequence of deformations of a superimposed grid. The method is applied by working backward from the final shape. The organ is subdivided into a grid of regions interconnected by 'springs' with resting lengths initially set according to the shape of the mature organ. Growth parameters just before cell divisions stop are then calculated for each region, on the basis of comparisons of clones induced late in development (Fig. 1b–e). The resting lengths of the springs are then adjusted according to these parameters, allowing each region to shrink, as the springs 'relax', to generate a revised grid and the

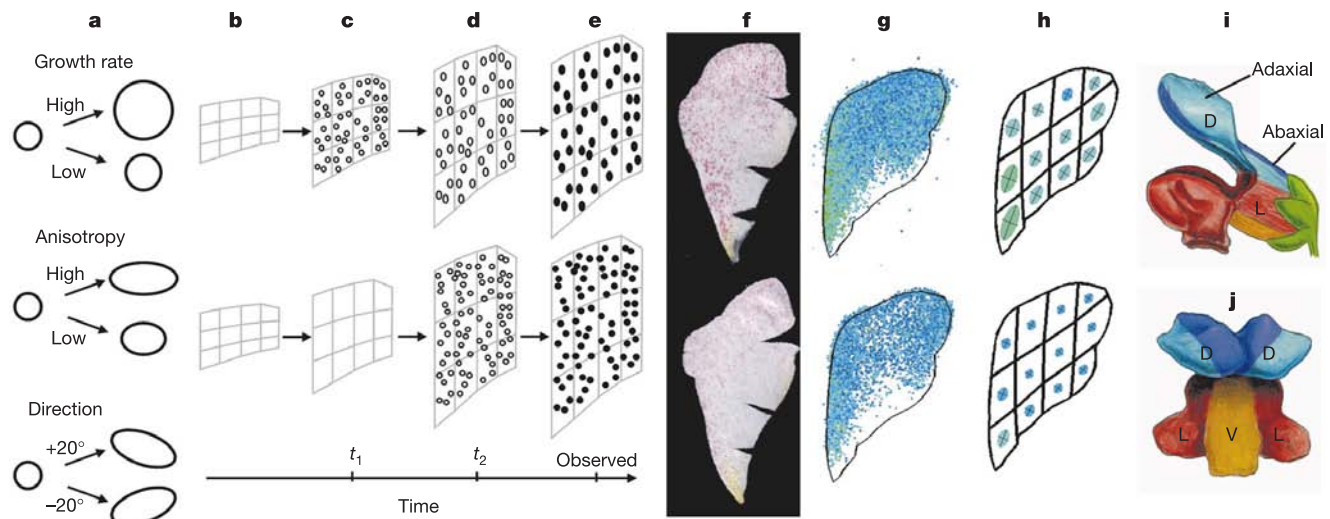


Figure 1 Determining growth parameters. **a**, Growth parameters responsible for regional shape changes, illustrated by the deformations of a growing circle. Growth rate affects area, anisotropy affects stretch and direction determines the angle of the stretch. **b–e**, Estimating growth parameters from clones. Clones are induced in a growing organ at two different developmental times, t_1 (top) and t_2 (bottom), in parallel experiments. By comparing the average characteristics (areas, shapes, directions) of the clones resulting from the two experiments (that is, by comparing the top and bottom of **e**), growth parameters for each region can be estimated for the (t_1, t_2) time interval. **f–h**, Gathering clonal data from *Antirrhinum* petals. Top and bottom show information from clones induced at P44 and P46 respectively. Clones were generated using an unstable *PALLIDA* (*PAL*) allele that carries a temperature-sensitive transposon²⁹. Transposon excision can be induced at specific times in a proportion of cells by switching plants grown at 25 °C to

15 °C for a short period³⁰. This treatment has no obvious effect on petal shape. Excision restores *PAL* gene function, resulting in clones of red epidermal cells in mature petals (**f**). For each induction time, clones from 23 flattened dorsal petal lobes on average were imaged and mapped onto an average petal shape (**g**) (A.-G. R.-L., S. J. Impey, E.C. and J.A.B., manuscript in preparation). Colour reflects clone size (blue is smaller than green). Average clone characteristics for different regions were computed for inductive time points separated by two plastochrons (**h**). For clarity, ellipse areas are scaled up by a factor of 20 relative to the petal size. Data from abaxial and adaxial lobe surfaces were treated separately. **i–j**, *Antirrhinum* flower, side (**i**), and face (**j**) view. Petals are colour-coded blue for dorsal (D), red for lateral (L) and yellow for ventral (V). The most dorsal half of the dorsal petal is shown in darker blue.

shape of the organ shortly before growth arrest. By repeating this procedure for earlier and earlier intervals, while aligning growth directions with respect to the revised grids, corresponding organ shapes and grids can be computed for earlier stages. The method assumes that clones are not dispersed during growth by cell movement, a reasonable assumption in the case of plants, where cells are usually fixed relative to each other.

This method was applied to petal development in *Antirrhinum*, a

well-characterized molecular genetic system for which a key question is how petal asymmetry emerges^{16–19}. *Antirrhinum* flowers have five petals, which are united in their proximal regions to form a tube while their distal regions form five distinct lobes (Fig. 1i, j). Dorsal and lateral petals are asymmetric, whereas the ventral petal is bilaterally symmetric. Petals are about seven cells thick, and the epidermal layer has been shown to make a major contribution to overall shape^{20–22}. Most epidermal cell divisions give daughter cells in the same layer²³. Floral meristems are generated from the apex at regular time intervals (10 h in this study), termed plastochrons²⁴. Petal primordia first emerge around P18 (plastochron 18) and reach maturity at P57 (Vincent, C. *et al.*, manuscript in preparation). Cell divisions mostly stop after P46 and further growth, resulting from cell expansion, has little impact on changes in petal shape (A.-G.R.-L. *et al.*, manuscript in preparation). In this study we computed growth of the dorsal petal lobe from P32 to P46 on the basis of the analysis of epidermal clones at maturity (Fig. 1f–h and Fig. 2a–c).

The growth rate was relatively constant, with an average doubling time of 21 h over about seven rounds of cell division. The average anisotropy was 1.15 (that is, 15% more cell divisions along the principal than in the minor direction of growth per doubling time). Although it is close to isotropic growth (1.0), the anisotropy is continuously maintained and therefore accumulates to generate a stretch in overall shape. The main growth direction was relatively uniform between regions at each stage but gradually rotated during development relative to the base of the lobe. The resulting deformation meant that clones growing at -76° (positive = clockwise) to the base of the lobe at P32 ended up with an orientation of -35° at maturity.

The results were validated in various ways. First, similar changes in shape were obtained for abaxial and adaxial surfaces, representing two independent data sets (Fig. 2b, c). Second, comparison of the shape inferred from the model with SEM (scanning electron micrographs) of a P32 lobe showed that the overall shapes were similar (Fig. 2d). Third, the changes in petal area calculated from the model were consistent with those obtained by directly measuring areas (Fig. 2e).

The pattern of cell shapes at P32 was also examined, using three-dimensional (3D) reconstructions to account for any effects of curvature⁶ (Fig. 2f). Cells were mainly elongated perpendicular to the base of the lobes (Fig. 2g). Cell wall orientations, which reflect the history of cell division planes²⁵, were most frequently perpendicular or at right angles to the base of the lobe (Fig. 2g). Thus, cells were axialized in a direction similar to the growth direction at P32 inferred from clonal analysis (Fig. 2b). This suggests that the directionality of growth depends on cell populations expanding and dividing preferentially in a given direction. By contrast, cell divisions in *Drosophila* wing development are thought to be

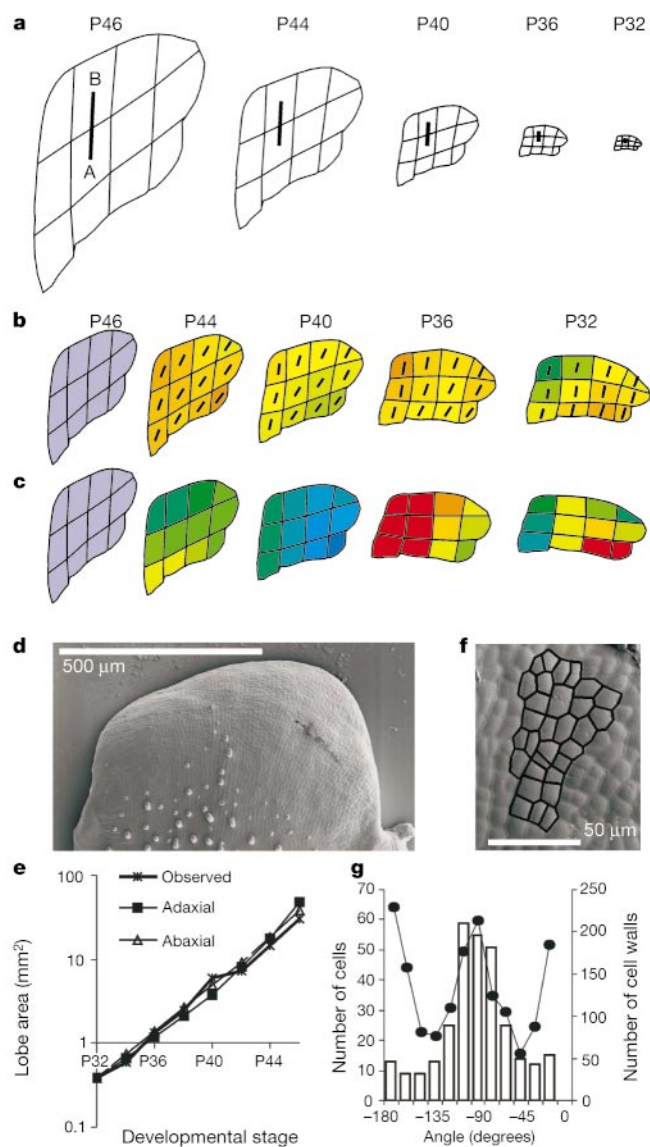


Figure 2 Results and validation of growth analysis. **a**, Changes in the shape, size and grid of the dorsal petal lobe from P46 to P32 based on changes in clones (not all stages are shown). Growth directions are computed relative to AB. **b**, **c**, Shapes as in **a** but scaled to the same size showing results for abaxial (**b**) and adaxial (**c**) data. The mature shape (P46) is shown in lilac. In **b**, the main growth direction is shown at the centre of each region and colour-coding refers to growth rate, which ranges from a doubling time of 15 h (orange) to 45 h (dark blue). In **c**, colour-coding refers to anisotropy (ratio of the increase along the main growth direction to the increase along the direction perpendicular to this, each time cell number doubles), which ranges from 1.04 (blue) to 1.5 (≥ 1.3 is red). **d**, Scanning electron micrograph (SEM) of a dorsal petal at P32. Scale bar, 500 μ m. **e**, Change in area from P32 to P46 based on the growth model compared to that based on direct measurements of petal lobes. **f**, Cell outlines on an SEM of a petal lobe at P32. Scale bar, 50 μ m. **g**, Distribution of angles of the main cell axis (bars) or cell walls (closed circles) relative to the base of the lobe, based on ten SEMs.

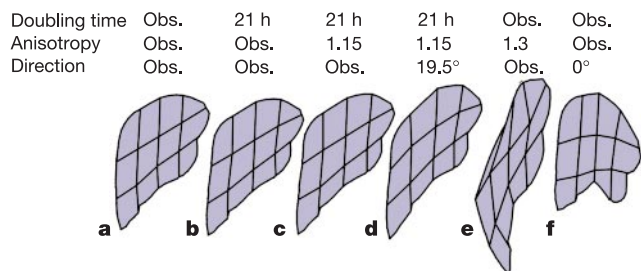


Figure 3 Petal lobe shapes at P46 resulting from growth simulations run forwards in developmental time from P32. **a**, Simulation incorporates observed (Obs.) growth parameters. **b**, Growth rate is set to its average over space and time. **c**, Growth rate and anisotropy are set to their average. **d**, All three growth parameters are set to their average. **e**, Anisotropy is set to 1.3. **f**, Growth direction is set parallel to AB in Fig. 2a.

randomly oriented, with directionality arising postmitotically¹⁴.

The contributions of growth parameters to final shape were explored by running the model forward in developmental time (Fig. 3a). To determine whether the asymmetry in petal shape reflects differential growth rates, a simulation was run with growth rate set to its average over space and time. This generated an asymmetric final petal shape roughly similar to the mature petal, showing that heterogeneity in growth rates was not critical (Fig. 3b). Further simulations showed that the asymmetry of the petal depended largely on the overall direction of anisotropic growth (Fig. 3c, d). The amount of anisotropy determined the extent to which the petal was stretched while the principal growth direction determined petal asymmetry (Fig. 3e, f). These transformations can be compared to oblique stretches of the kind proposed by Thompson²⁶ to describe the relationship between different biological forms, although he used them mainly as a static measure rather than in the dynamic developmental sense used here.

The base of the lobe, which corresponds to the boundary between lobe and tube (Fig. 4a), turns through an angle of about 45° during development as the tube grows more on its dorsal side (Fig. 4b–d). If this is taken into account and the petal lobes are oriented relative to the flower as a whole, it becomes evident that the growth direction is maintained roughly parallel to the proximodistal axis throughout development (Fig. 4e). This suggests that a long-range signal

(arrows in Fig. 4b–d) acts continuously during development to maintain the growth direction along the proximodistal axis of the petal as a whole. Long-range signals have been described before with regard to patterns^{27,28} but this is the first time one has been linked quantitatively to directional growth and the generation of shape.

On the basis of this analysis, the main transformations in petal lobe shape can be captured by the following model. Growth rates and anisotropy are fixed at 21 h and 1.15 respectively. The long-range signal is modelled by assuming that the direction of growth is fixed (vertical in Fig. 4f) while the lobe rotates through an angle of 45°. This generates a good approximation of the final shape with minimum parameter complexity.

Received 13 August 2002; accepted 20 January 2003; doi:10.1038/nature01443.

1. Avery, G. S. *Structure and development of the tobacco leaf*. *Am. J. Bot.* **20**, 565–592 (1933).
2. Richards, O. W. & Kavanagh, A. J. The analysis of the relative growth gradients and changing form of growing organisms: Illustrated by the tobacco leaf. *Am. Nat.* **77**, 385–399 (1943).
3. Erickson, R. O. Modeling of plant growth. *Annu. Rev. Plant Physiol.* **27**, 407–434 (1976).
4. Erickson, R. O. Relative elemental rates and anisotropy of growth in area: a computer programme. *J. Exp. Bot.* **17**, 390–403 (1966).
5. Wolf, S. D., Silk, W. K. & Plant, R. E. Quantitative patterns of leaf expansion—Comparison of normal and malformed leaf growth in *Vitis-Vinifera* Cv Ruby Red. *Am. J. Bot.* **73**, 832–846 (June 1986).
6. Dumais, J. & Kwiatkowska, D. Analysis of surface growth in shoot apices. *Plant J.* **31**, 229–241 (2002).
7. Hernandez, F., Havelange, A., Bernier, G. & Green, P. B. Growth behavior of single epidermal cells during flower formation: Sequential scanning electron micrographs provide kinematic patterns for *Anagallis*. *Planta* **185**, 139–147 (1991).
8. Subtelny, S. & Sussex, I. M. (eds) *The Clonal Basis of Development* (Academic, New York, 1978).
9. Poethig, R. S. & Sussex, I. M. The cellular parameters of leaf development in tobacco: a clonal analysis. *Planta* **165**, 170–184 (1985).
10. Poethig, R. S. & Szymkowiak, R. J. Clonal analysis of leaf development in maize. *Maydica* **40**, 67–76 (1995).
11. Dolan, L. & Poethig, R. S. Clonal analysis of leaf development in cotton. *Am. J. Bot.* **85**, 315–321 (1998).
12. Garcia-Bellido, A. & Merriam, J. R. Parameters of the wing imaginal disc development of *Drosophila melanogaster*. *Dev. Biol.* **24**, 61–87 (1971).
13. Gonzalez-Gaitan, M., Capdevila, M. P. & Garcia-Bellido, A. Cell proliferation patterns in the wing imaginal disc of *Drosophila*. *Mech. Dev.* **40**, 183–200 (1994).
14. Resino, J., Salama-Cohen, P. & Garcia-Bellido, A. Determining the role of patterned cell proliferation in the shape and size of the *Drosophila* wing. *Proc Natl Acad. Sci. USA* **99**, 7502–7507 (2002).
15. Poethig, R. S. Clonal analysis of cell lineage patterns in plant development. *Am. J. Bot.* **74**, 581–594 (1987).
16. Almeida, J., Rocheta, M. & Galego, L. Genetic control of flower shape in *Antirrhinum majus*. *Development* **124**, 1387–1392 (1997).
17. Luo, D. *et al.* Control of organ asymmetry in flowers of *Antirrhinum*. *Cell* **99**, 367–376 (1999).
18. Luo, D., Carpenter, R., Vincent, C., Copsey, L. & Coen, E. Origin of floral asymmetry in *Antirrhinum*. *Nature* **383**, 794–799 (1995).
19. Cubas, P., Lauter, N., Doebley, J. & Coen, E. The TCP domain: a motif found in proteins regulating plant growth and development. *Plant J.* **18**, 215–222 (1999).
20. Vincent, C. A., Carpenter, C. & Coen, E. S. *Plant J.* **33**, 1–10 (2003).
21. Perbal, M. C., Haughn, G., Saedler, H. & Schwarz-Sommer, Z. Non-cell-autonomous function of the *Antirrhinum* floral homeotic proteins DEFICIENS and GLOBOSA is exerted by their polar cell-to-cell trafficking. *Development* **122**, 3433–3441 (1996).
22. Efremova, N. *et al.* Epidermal control of floral organ identity by class B homeotic genes in *Antirrhinum* and *Arabidopsis*. *Development* **14**, 2661–2671 (2001).
23. Tilney-Basset, R. A. E. *Plant Chimeras* (Edward Arnolds, London, 1986).
24. Maksymowych, R. *Analysis of Leaf Development* Developmental and cell biology series 1 (Cambridge Univ. Press, London, 1973).
25. Lloyd, C. W. & Traas, J. A. The role of F-actin in determining the division plane of carrot suspension cells—drug studies. *Development* **102**, 211–221 (1988).
26. Thompson, D. A. W. *On Growth and Form*, 2nd edn Vol. 2 (Cambridge Univ. Press, Cambridge, UK, 1942).
27. Sachs, T. in *Pattern Formation in Plant Tissues* (eds Barlow, P. W., Bray, D., Green, P. B. & Slack, J. M. W.) (Cambridge Univ. Press, Cambridge, UK, 1991).
28. Teleman, A., Strigini, M. & Cohen, S. M. Shaping morphogen gradients. *Cell* **105**, 559–562 (2001).
29. Vincent, C. A., Carpenter, R. & Coen, E. S. Cell lineage patterns and homeotic gene activity during *Antirrhinum* flower development. *Curr. Biol.* **5**, 1449–1457 (1995).
30. Harrison, B. J. & Fincham, J. R. S. Instability at the *Pal* locus in *Antirrhinum Majus*. I. Effects of environment on frequencies of somatic and germinal mutation. *Heredity* **19**, 237–258 (1964).

Acknowledgements We thank R. Carpenter for providing plant stocks, C. Vincent and K. Lee for the scanning electron microscopy, J. Dumais for help with SEM 3D reconstructions, N. Orme for drawings in Fig. 1, S. J. Impey for gathering the image database and for early work on the project, supported by a BBSRC grant. A.-G. R.-L. was supported by a Norwich Research Park studentship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.C. (e-mail: Enrico.Coen@bbsrc.ac.uk).

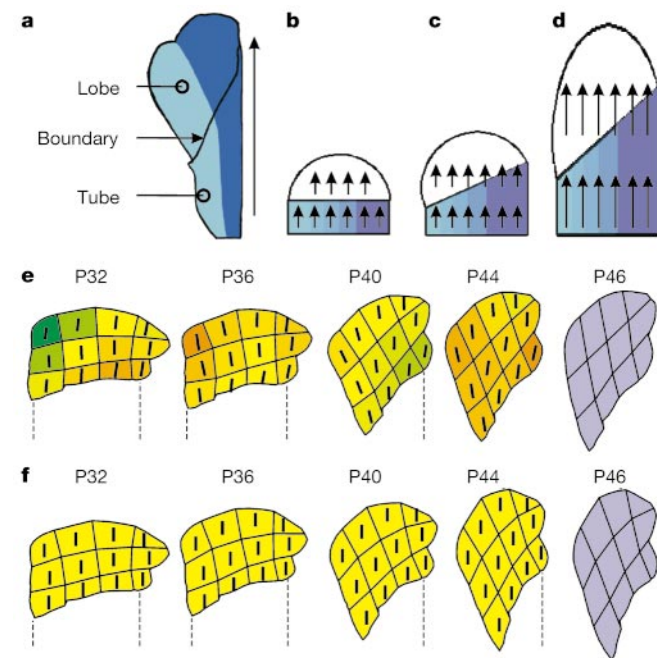


Figure 4 Growth of petal lobe in relation to the whole flower. **a**, Flattened dorsal petal colour-coded as in Fig. 1i, j, showing the boundary between tube and lobe. Arrow indicates the proximodistal axis. **b–d**, Diagram showing the change in shape from P32 to maturity when the growth direction is continuously coordinated along the whole petal by a long-range signal (arrows, arbitrarily shown pointing up rather than down). Lobe is white and tube is blue, with dorsal regions in darker blue. Initially, the lobe is bilaterally symmetrical and the tube–lobe boundary is perpendicular to the proximodistal axis of the petal (**b**). The dorsal side of the tube grows preferentially, resulting in a change in the orientation of the tube–lobe boundary (**c**). As the growth direction is maintained parallel to the proximodistal axis, anisotropic growth results in the lobe becoming asymmetric (**d**). **e**, Observed growth directions are mainly parallel to the proximodistal axis when the shapes are oriented relative to the tube (shown as dotted lines). **f**, Minimal model. Growth rates and anisotropy are fixed at 21 h and 1.15 respectively. The long-range signal is modelled by assuming that the growth direction is fixed while the lobe rotates through an angle of 45°.

Selection of evolutionarily conserved mucosal-associated invariant T cells by MR1

Emmanuel Treiner^{*,†}, Livine Duban^{*,†}, Seiamak Bahram[‡], Mirjana Radosavljevic[‡], Valerie Wanner[‡], Florence Tilloy^{*}, Pierre Affaticati^{*}, Susan Gilfillan[§] & Olivier Lantz^{*}

^{*} Laboratoire d'Immunologie and INSERM U520, Institut Curie, Paris 75005, France

[‡] INSERM-CReS, Centre de Recherche d'Immunologie et d'Hématologie, 4 rue Kirschleger, 67085 Strasbourg Cedex, France

[§] Department of Pathology and Immunology, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA

[†] These authors contributed equally to this work

The evolutionary conservation of T lymphocyte subsets bearing T-cell receptors (TCRs) using invariant α -chains is indicative of unique functions. CD1d-restricted natural killer T (NK-T) cells that express an invariant V α 14 TCR α chain have been implicated in microbial and tumour responses, as well as in auto-immunity^{1,2}. Here we show that T cells that express the canonical hV α 7.2-J α 33 or mV α 19-J α 33 TCR rearrangement³ are preferentially located in the gut lamina propria of humans and mice, respectively, and are therefore genuine mucosal-associated invariant T (MAIT) cells. Selection and/or expansion of this population requires B lymphocytes, as MAIT cells are absent in B-cell-deficient patients and mice. In addition, we show that MAIT cells are selected and/or restricted by MR1, a monomorphic major histocompatibility complex class I-related molecule that is markedly conserved in diverse mammalian species⁴. MAIT cells are not present in germ-free mice, indicating that commensal flora is required for their expansion in the gut lamina propria. This indicates that MAIT cells are probably involved in the host response at the site of pathogen entry, and may regulate intestinal B-cell activity.

NK-T and MAIT cells are defined by their use of a restricted TCR repertoire: an invariant TCR α chain—one V α -J α combination—with a CDR3 of constant length paired with a limited number of V β segments^{1–3}. T cells that express an invariant TCR α chain have an activated phenotype (CD44^{hi}), and their high frequency in particular lymphocyte subsets and organs results from oligoclonal expansion. They can be CD4⁺ (CD4 positive), CD4⁺ CD8[–] (double negative or DN) or CD8 α in humans, and do not seem to require CD4 or CD8 accessory molecules for selection. NK-T and MAIT cells are restricted by β 2-microglobulin (β 2m)-dependent major histocompatibility complex (MHC) class Ib molecules in a TAP (transporters associated with antigen processing)-independent manner: CD1d for NK-T cells and a distinct β 2m-dependent molecule for MAIT cells. NK-T cells are abundant in the thymus, liver, spleen and bone marrow; however, the exact tissue distribution of T cells that express hV α 7.2/mV α 19-J α 33 is not known³.

To assess the distribution of MAIT cells, we used fluorescence-activated cell sorting (FACS) to separate TCR $\alpha\beta$ DN T cells from human peripheral blood lymphocytes (PBL) according to the expression level of the gut-specific α 4 β 7 (ref. 5) homing marker, and quantified hV α 7.2-J α 33 transcripts in different fractions. The invariant transcripts were enriched in the TCR $\alpha\beta$ DN α 4 β 7⁺ fraction, as shown by a six-cycle shift to the left of the polymerase chain reaction (PCR) curve, suggesting that hV α 7.2-J α 33⁺ T cells are associated with the gut (Fig. 1a). Enrichment of hV α 7.2-J α 33 transcripts in 11 gut biopsies, as compared with three skin samples, in assays using DN T cells as a positive control and CD4⁺ PBL as a negative control verified the gut location of hV α 7.2-J α 33⁺ T cells (Fig. 1b). Gut-associated T cells comprise intra-epithelial, Peyer's

patch and lamina propria (LPL) lymphocytes⁶. A preliminary analysis of gut biopsies containing variable amounts of intra-epithelial, Peyer's patch and lamina propria tissues suggested that the hV α 7.2-J α 33⁺ T cells were restricted to the lamina propria (data not shown).

To determine whether this gut-specific homing pattern also occurs in mice, we quantified canonical mV α 19-J α 33 expression in gut lymphocytes of B6 mice and in those of TAP- and invariant chain (Ii)-deficient (TAP^{–/–}Ii^{–/–}) mice. Rare in wild-type mice, mV α 19-J α 33⁺ T cells are more abundant in TAP^{–/–}Ii^{–/–} mice, in which the numbers of mainstream T cells are reduced owing to low cell-surface expression of MHC class Ia and II molecules^{3,7}. mV α 19-J α 33 transcripts were tenfold more abundant in LPL than in the mesenteric lymph nodes (MLN) of B6 mice; however, they were absent in LPL of β 2m^{–/–} mice (Figs 1c and 2e). In addition, tenfold more mV α 19-J α 33 transcripts were amplified from TAP^{–/–}Ii^{–/–} LPL than from B6 LPL, confirming that the selection/expansion of MAIT cells is independent of TAP and Ii. mV α 19-J α 33⁺ transcripts were abundant in lamina propria and Peyer's patch samples (albeit less so in Peyer's patch), but not in intra-epithelial lymphocytes (Fig. 1d). Polyclonal sequencing of mV α 19-J α 33 amplicons confirmed the over-representation of the canonical CDR3 specific for this invariant TCR α chain in the gut lamina propria (Fig. 2e). Together, these results show that the gut lamina propria is a favoured location for β 2m-dependent MAIT cells in humans and mice. In both species, preliminary data indicate that hV α 7.2/mV α 19-J α 33⁺ T cells are also present in the lung (data not

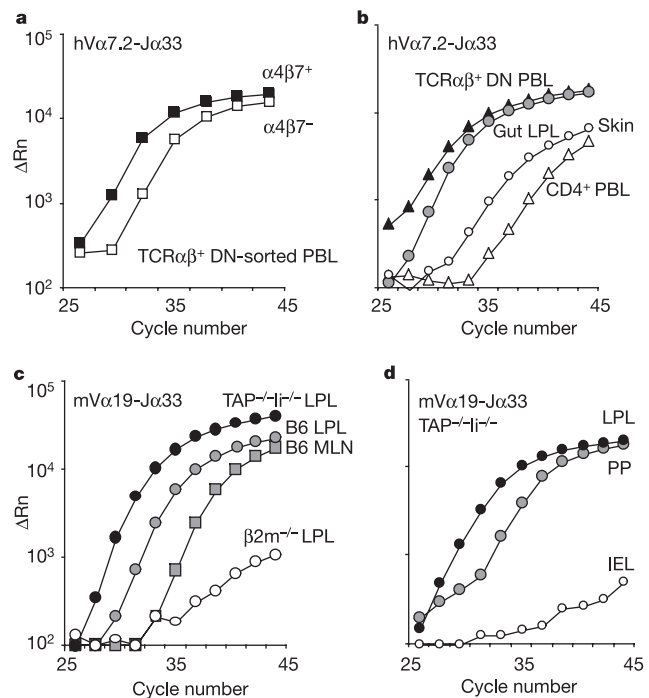


Figure 1 MAIT cells are located in the gut lamina propria of both humans and mice. Lymphocytes were isolated from the indicated organs and RNA was extracted. After cDNA synthesis, kinetic Q-PCR was carried out for C α and the indicated V α -J α combination. Data shown are representative of at least three independent experiments (two for panel a). **a**, hV α 7.2-J α 33⁺ T cells are enriched in the α 4 β 7⁺ fraction of TCR $\alpha\beta$ DN human PBL. α 4 β 7⁺ and α 4 β 7[–] fractions from CD4⁺ CD8[–] TCR β ⁺ PBL were isolated by FACS. **b**, hV α 7.2-J α 33⁺ T cells are present in the human gut. DN and CD4 TCR β ⁺ T cells were isolated from PBL. RNA from skin or gut was isolated from frozen biopsies. Data are representative of 11 (gut) and three (skin) samples. **c**, mV α 19-J α 33⁺ T cells are present in the gut lamina propria and MLN in mice; they are more abundant in TAP^{–/–}Ii^{–/–} mice. **d**, mV α 19-J α 33⁺ T cells are present in the gut Peyer's patch (PP) and in LPL, but not in intra-epithelial lymphocytes (IEL). Δ Rn, change in normalized fluorescence ratio; PBL, peripheral blood lymphocytes.

shown), suggesting that MAIT cells may be present in several mucosal tissues.

Mainstream T cells are selected by classical MHC class I and class II molecules expressed on thymic epithelium⁸, whereas NK-T cells are selected by recognition of CD1d expressed on haematopoietic cells, specifically CD4⁺CD8⁺ (double positive or DP) thymocytes⁹. Urdahl *et al.* recently proposed that most MHC class Ib-restricted T cells may rely on haematopoietic cells for selection¹⁰. As MAIT cells require a $\beta 2m$ -dependent molecule for selection³, we generated fetal liver chimaeras between $\beta 2m$ -deficient and wild-type B6 mice to determine which cell type mediates their selection. As expected, B6 $\rightarrow \beta 2m^{-/-}$ (B6 fetal liver cells transferred into an irradiated $\beta 2m$ -deficient recipient) chimaeras have very few CD8⁺ T cells⁸ (see Supplementary Fig. S1). Conclusively, mV $\alpha 19$ -J $\alpha 33$ transcripts were present in B6 $\rightarrow \beta 2m^{-/-}$ but not in $\beta 2m^{-/-} \rightarrow$ B6 chimaeras (Fig. 2a). Thus, expression of a $\beta 2m$ -dependent MHC class Ib molecule on haematopoietic cells is necessary and sufficient for the selection/expansion of DN MAIT cells.

To identify the haematopoietic cell type responsible for the selection of MAIT cells, we generated mixed chimaeras in which the bone marrow from $\beta 2m$ -deficient mice (BM-1) was mixed with several sources of bone marrow incapable of generating T cells (BM-2), and then injected into $\beta 2m$ -deficient lymphoid hosts ($\beta 2m^{-/-}$ RAG2^{-/-} $\gamma c^{-/-}$; that is, hosts lacking $\beta 2m$, recombination-activating gene 2 and the common interleukin receptor γ -chain, respectively). In this setting, T cells repopulating the host thymus are derived from BM-1 and are $\beta 2m$ -deficient, whereas cells originating from BM-2 are the only source of $\beta 2m$. Therefore, when BM-2 is RAG2^{-/-}, only macrophages and dendritic cells express $\beta 2m$ -dependent MHC molecules, whereas when BM-2 is CD3 $\epsilon^{-/-}$, $\beta 2m^{+}$ B cells are also present. In all groups, very few CD8^{hi} T cells

were generated (see Supplementary Fig. S2). When only macrophages and dendritic cells were $\beta 2m^{+}$, no mV $\alpha 19$ -J $\alpha 33$ transcripts could be detected in LPL. By contrast, the addition of $\beta 2m^{+}$ B cells was sufficient to give rise to mV $\alpha 19$ -J $\alpha 33^{+}$ LPL (Fig. 2b). These results indicate that B cells mediate MAIT cell selection/expansion, whereas dendritic cells and macrophages do not. The selection of MAIT cells by B cells was confirmed by a significant reduction in mV $\alpha 19$ -J $\alpha 33$ transcripts in TCR $\alpha\beta^{+}$ DN cells from the MLN and LPL of mice negative for the immunoglobulin heavy chain joining region ($J_H^{-/-}$), mice, which are completely B-cell-deficient¹¹ (Fig. 2c and data not shown). Mice without the transmembrane region of immunoglobulin- μ (IgM; $\mu MT^{-/-}$ mice) lack most B cells, but do harbour IgA-producing B cells in their intestine^{12,13}. Normal numbers of MAIT cells were found in LPL from $\mu MT^{-/-}$ mice, showing that these residual gut B cells are sufficient for the accumulation of MAIT cells (Fig. 2c). Finally, mV $\alpha 19$ -J $\alpha 33$ transcripts were amplified equally well from the lamina propria and MLN of xid and control mice (data not shown), indicating that B1 B cells, absent in xid mice¹⁴, are not involved in the selection/expansion of MAIT cells.

To establish whether human MAIT cells are also dependent on B cells for selection/expansion, we quantified hV $\alpha 7.2$ -J $\alpha 33$ transcripts in TCR $\alpha\beta^{+}$ DN human PBL from normal individuals and four patients with mutated Bruton tyrosine kinase¹⁵. hV $\alpha 7.2$ -J $\alpha 33$ transcripts in DN T cells were extremely rare in two patients, diminished tenfold in another, and unaffected in the fourth, as compared with all normal subjects tested so far (Fig. 2d). Analogous to the situation in μMT -deficient mice, the normal levels of hV $\alpha 7.2$ -J $\alpha 33$ transcripts observed in one patient could be related to the presence of some B cells in his gut, despite the absence of B cells (CD19⁺) in his blood (data not shown). These results show that B cells are necessary for the selection/expansion of MAIT cells in both

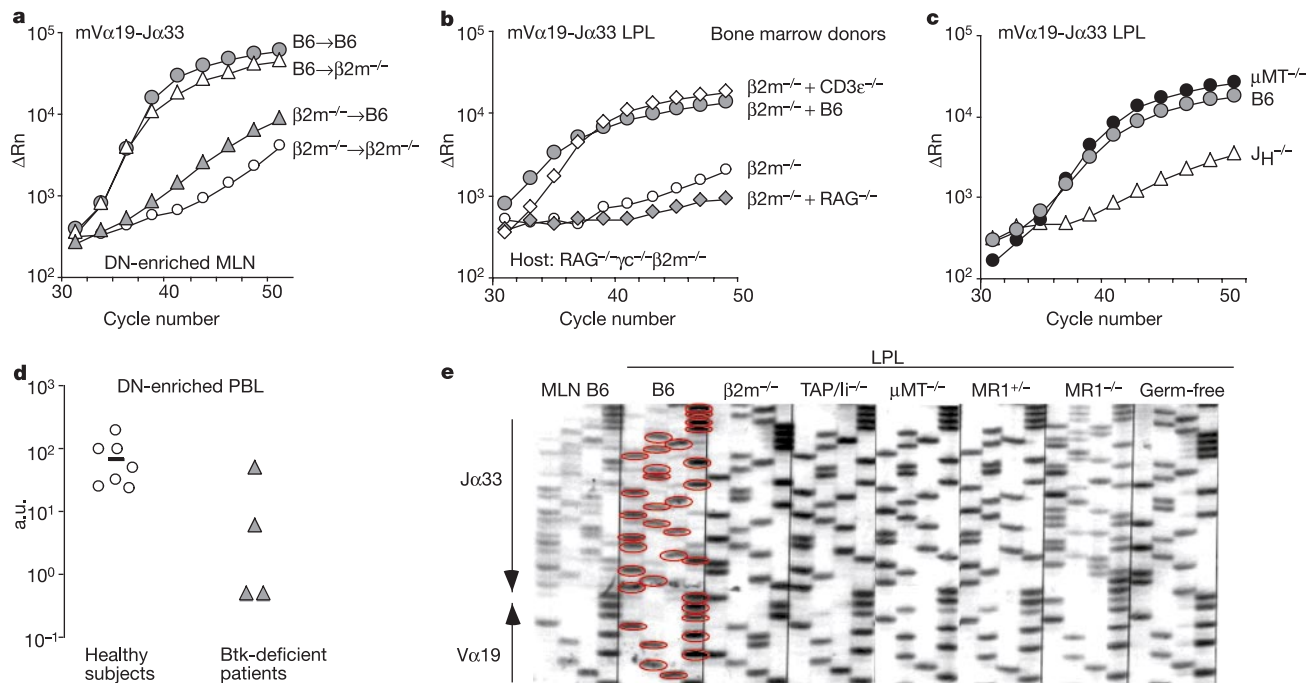


Figure 2 MAIT cells are selected by B cells. **a**, mV $\alpha 19$ -J $\alpha 33^{+}$ T cells are selected by haematopoietic cells. Fetal liver chimaeras with the indicated genotype were prepared as described in the Methods. DN-enriched fractions from MLN were isolated and Q-PCR was carried out. These data are representative of at least three samples per group analysed in two independent experiments. **b**, mV $\alpha 19$ -J $\alpha 33^{+}$ T cells are selected by B cells. 5-Fluorouracil-treated RAG $^{-/-}$ $\gamma c^{-/-}$ $\beta 2m^{-/-}$ recipients were reconstituted with a 1:1 mixture of the indicated bone marrow. After three months, LPL were isolated and Q-PCR was carried out. These data are representative of at least two samples per group. **c**, mV $\alpha 19$ -J $\alpha 33^{+}$ T cells are selected by B cells. LPL were isolated from the indicated

mice and Q-PCR was carried out. Seven mice in each group were analysed; representative data are shown. **d**, hV $\alpha 7.2$ -J $\alpha 33^{+}$ T cells are selected by human B cells. DN-enriched T cells were obtained from the indicated subjects and Q-PCR was performed. Results are normalized to C_{α} and expressed as a percentage relative to one sample, to which we attributed a value of 100 arbitrary units (a.u.). Btk, Bruton's tyrosine kinase. **e**, The mV $\alpha 19$ -J $\alpha 33$ canonical sequence is not found in MR1-deficient, $\beta 2m$ -deficient or germ-free mice. V $\alpha 19$ -J $\alpha 33$ amplification was carried out on the indicated samples and polyclonal amplicons sequenced with a V $\alpha 19$ primer. The canonical sequence is circled in red in the B6 LPL sample.

humans and mice, and suggest that MAIT cells interact with lamina propria B cells.

Thus, B cells express a $\beta 2m$ -dependent, TAP- and Ii-independent MHC class Ib molecule distinct from CD1d that selects MAIT cells. MR1 is by far the most highly conserved MHC class I-related molecule in mammalian species^{16–18}: the 90% sequence identity between mouse and human is unparalleled^{4,19}. Like CD1d, MR1 is encoded on human chromosome 1, but is more closely related to classical class I molecules than are other class Ib family members. These attributes made MR1 a plausible candidate ligand for the MAIT TCR. To demonstrate recognition of MR1 by mV $\alpha 19$ -J $\alpha 33$ ⁺ T cells, we generated several cell lines expressing a full-length murine MR1 complementary DNA fused to enhanced green fluorescent protein (EGFP; see Supplementary Fig. S3), and examined their ability to stimulate a series of mV $\alpha 19$ -J $\alpha 33$ ⁺ T–T hybridomas (ref. 3 and data not shown). Three (18G7, 6C2 and 6H2) out of more than 20 tested hybridomas were specifically stimulated to release interleukin 2 (IL-2) by the MR1 transfectants (Fig. 3a), and anti-TCR antibodies blocked this stimulation (Fig. 3a and

Supplementary Fig. S3c). These results strongly suggest that mV $\alpha 19$ -J $\alpha 33$ ⁺ T cells directly recognize MR1.

To assess the role of MR1 in the selection and expansion of MAIT cells *in vivo*, we analysed mice in which the $\alpha 1$ and $\alpha 2$ domains of MR1 were deleted using gene targeting (see Supplementary Fig. S4). MR1^{−/−} mice contained phenotypically normal T, B and NK cells at the expected number and frequency in the central and peripheral lymphoid organs (thymus, spleen, peripheral lymph nodes (PLN) and MLN) (see Supplementary Fig. S5). mV $\alpha 19$ -J $\alpha 33$ sequences were readily amplified from MLN or lamina propria T cells obtained from littermate controls, but not in those from MR1^{−/−} mice (Fig. 3b and data not shown). These results were confirmed by polyclonal sequencing of mV $\alpha 19$ -J $\alpha 33$ amplicons obtained from lamina propria cells (Fig. 2e). To evaluate the frequency of the canonical sequence in MR1-deficient mice, we sequenced cloned mV $\alpha 19$ -J $\alpha 33$ rearrangements amplified from genomic LN DNA. As shown in Table 1, 80% of the mV $\alpha 19$ -J $\alpha 33$ sequences from MR1⁺ mice corresponded to the invariant chain, and only 6% of the sequences were out of frame, demonstrating a very strong selection for this canonical sequence. In sharp contrast to this, only 7% of mV $\alpha 19$ -J $\alpha 33$ junctions sequenced from MR1-deficient mice encoded the invariant TCR α chain, and 47% of the sequences obtained were out of frame. The presence of a few canonical sequences in MR1^{−/−} (and $\beta 2m$ ^{−/−}) mice reflects a strong mechanistic bias for this particular rearrangement, which is germ-line encoded and favoured by homology-directed recombination^{20,21}. mV $\alpha 19$ -J $\alpha 33$ chains with the same CDR3 length as the invariant

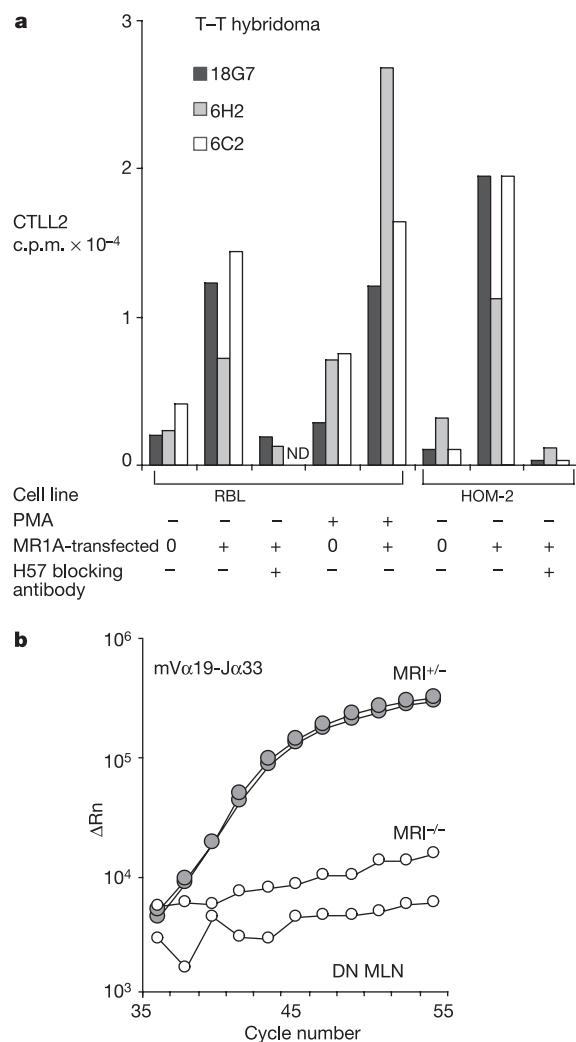


Figure 3 MR1 is the restricting MHC molecule for MAIT cells. **a**, IL-2 production by T–T hybridomas stimulated by MR1 transfectants. A panel of mV $\alpha 19$ T–T hybridomas was stimulated with the indicated cell lines with or without anti-TCR β monoclonal antibody (1 μ g ml^{−1}) for 48 h. Supernatants were assayed using the IL-2-dependent cell line CTLL2. PMA, phorbol myristate acetate. **b**, The mV $\alpha 19$ -J $\alpha 33$ junction is not present in DN T cells from MR1^{−/−} MLN. DN TCR $\alpha\beta$ ⁺ T cells were isolated from the MLN of MR1^{−/−} and MR1^{+/−} mice, and Q-PCR performed with C α and mV $\alpha 19$ -J $\alpha 33$ specific primers. ND, not determined.

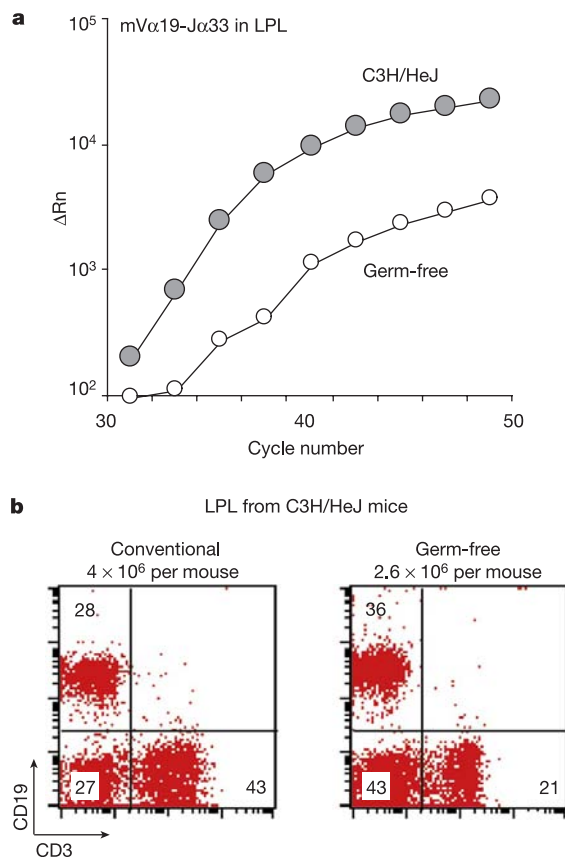


Figure 4 MAIT cells are absent in germ-free mice. **a**, Absence of MAIT cells in germ-free mice. LPL were isolated from germ-free or control C3H mice, and Q-PCR was carried out. Data shown are representative of seven samples per group studied in three independent experiments. **b**, Germ-free mice have significant numbers of CD19⁺ cells. LPL were isolated from the indicated mice, and stained with anti-CD19 and anti-CD3 ϵ monoclonal antibodies. Numbers in each quadrant indicate the percentage of each population.

TCR α chain were also enriched in MR1⁺ mice, some partially encoded by N nucleotides. Thus, mV α 19-J α 33 resembles the canonical mV γ 3-J γ 1 and mV δ 1-D1-D2-J δ 2 chains expressed by murine $\gamma\delta$ dendritic epidermal T cells (DETC), which are mechanistically favoured by homology-directed recombination as well as being selected for at the protein level^{21–23}. Overall, 92% of the MR1⁺ mV α 19-J α 33 sequences were either canonical or of identical CDR3 length, whereas only 12% of the MR1^{−/−} sequences fit these criteria. Together, these results show that MR1 is, indeed, the β 2m-dependent class Ib molecule that is required for the selection/expansion of T cells expressing the invariant mV α 19-J α 33 chain *in vivo*.

MAIT cells may bind MR1 alone or in association with a ligand. NK-T cells recognize as yet unidentified CD1d-bound endogenous glycolipids that are present in fetal thymus organ cultures¹. NK-T cell numbers and phenotype are also unchanged in germ-free mice²⁴. By contrast, we found that mV α 19-J α 33 transcripts were reduced in LPL (Fig. 4a) and in DN MLN T cells obtained from germ-free mice (data not shown). The canonical sequence was absent from LPL, as shown by polyclonal sequencing (Fig. 2e), despite the presence of a significant number of CD19⁺ B cells (Fig. 4b). Thus, the accumulation of MAIT cells is dependent on gut commensal flora and MR1. It is possible that MR1 remains

intracellular in the absence of stress and/or the appropriate ligand, as do MIC (MHC class I chain-related), H2-M3 and HLA-E (human leukocyte antigen E)^{25–27}. In support of this, MR1 molecules expressed by transfection remain largely in the endoplasmic reticulum, associated with molecular chaperones that have been implicated in MHC class Ia assembly. Furthermore, recombinant MR1 heavy chains appear to undergo a folding event similar to class Ia molecules when expressed in highly supplemented medium containing a putative ligand (M. J. Miley *et al.*, unpublished observations). In view of this, it is attractive to speculate that the cell-surface expression of MR1 requires a ligand, which may not be a peptide, derived from or induced by commensal flora. Alternatively, microbial products could facilitate translocation of MR1 to the cell membrane, where it is directly recognized by the hV α 7.2/mV α 19-J α 33 TCR.

NK-T cells and MAIT cells are two phylogenetically conserved MHC class Ib-restricted T-cell subpopulations that share many features: both use a semi-invariant TCR to recognize their restricting elements in a co-receptor-independent manner, and seem to display some level of autoreactivity *in vivo*^{1–3}. Here we show that they share a similar developmental pathway, as MAIT cells also require interaction with bone-marrow-derived cells for selection/expansion. NK-T cells bind CD1d on cortical DP thymocytes⁹,

Table 1 Sequences of V α 19-J α 33 amplicons from MR1-deficient mice and controls

AV19				Canonical junction				CATGGATAGCAACTATCAGTTGAT				AJ33			
TGTGCTGTGAGGGATCAGCGC				TGTGCTGTGAGGGATAGCAACTATCAGTTGAT											
MR1-deficient				β2m-deficient				MR1-positive littermate controls							
B	CVV	MDGNYQW	IWGS	B	CVV	MDGNYQW	IWGS	B	CVV	MDGNYQW	IWGS				
H	CAV	KDSNYQL	IWAG	H	CAV	KDSNYQL	IWAG	H	CAV	KDSNYQL	IWAG				
M	CAV	RDSNYQL	IWGS	M	CAV	RDSNYQL	IWGS	M	CAV	RDSNYQL	IWGS				
Mouse 1 lymph nodes				Lymph nodes				Mouse 1 lymph nodes							
CAV	RDSNYQL	IWGS	(2/11)	CAV	RDSNYQL	IWGS	(3/33)	CAV	RDSNYQL	IWGS	(9/15)				
CAV	RRPSNYQL	IWGS	(1/11)	CAV	RLGSNYQL	IWGS	(6/33)	CAV	MDSNYQL	IWGS	(2/15)				
Out of frame			(8/11)	CAV	RDRDSNYQL	IWGS	(15/33)	CAP	MDSNYQL	IWGS	(2/15)				
			(7 unique)	CAV	RAEGDSNYQL	IWGS	(3/33)	CAV	REGNYQL	IWGS	(1/15)				
				Out of frame			(6/33)	Out of frame			(1/15)				
							(3 unique)								
Mouse 2 lymph nodes								Mouse 2 lymph nodes							
CAA	HSNYQL	IWGS	(4/34)					CAV	RDSNYQL	IWGS	(18/26)				
CAV	REGYQL	IWGS	(1/34)					CAV	MDSNYQL	IWGS	(1/26)				
CAV	RDSNYQL	IWGS	(1/34)					CAV	LDSNYQL	IWGS	(2/26)				
CAV	MVDSNYQL	IWGS	(1/34)					CAV	RDRNYQL	IWGS	(1/26)				
CAV	RPDSNYQL	IWGS	(2/34)					CAV	RPMDSNYQL	IWGS	(1/26)				
CAV	RDRSNYQL	IWGS	(1/34)					Out of frame			(3/26)				
CAV	RDRDSNYQL	IWGS	(15/34)									(3 unique)			
CAV	RDGDSNYQL	IWGS	(4/34)												
Out of frame			(5/34)												
			(3 unique)												
Mouse 3/4 pooled DN-enriched								Mouse 3/4 pooled DN-enriched							
CAV	RDSNYQL	IWGS	(5/34)					CAV	RDSNYQL	IWGS	(23/26)				
CAV	RVDSNYQL	IWGS	(3/34)					CAV	MDSNYQL	IWGS	(1/26)				
CAV	RDRSNYQL	IWGS	(3/34)					CAV	LDSNYQL	IWGS	(1/26)				
CAV	RDPGDSNYQL	IWGS	(1/34)					Out of frame			(1/26)				
Out of frame			(22/34)												
			(6 unique)												
Mouse 5-7 pooled DN-enriched								Mouse 5-7 pooled DN-enriched							
CAV	RDRDYQL	IWGS	(5/30)					CAV	RDSNYQL	IWGS	(30/34)				
CAV	RVDSNYQL	IWGS	(3/30)					CAV	ADSNYQL	IWGS	(1/34)				
CAV	REGSNYQL	IWGS	(2/30)					CAV	RDRMDSNYQL	IWGS	(1/34)				
CAV	RGDSNYQL	IWGS	(1/30)					CAV	RDGRDSNYQL	IWGS	(1/34)				
CAV	RVGGNYQL	IWGS	(1/30)					Out of frame			(1/34)				
CAV	RDQGSNYQL	IWGS	(1/30)												
CAV	RDLGSNYQL	IWGS	(1/30)												
Out of frame			(16/30)												
			(10 unique)												

Generation of the canonical junction by the germ-line AV19 and AJ33 segments is shown, with the 4-bp stretch of shared homology in bold and P nucleotides underlined. The amino acid sequences of the bovine (B), human (H) and murine (M) canonical junctions are shown above each data set, and canonical junctions are underlined within the data sets. diff. seq., different sequence.

whereas MAIT cells probably recognize MR1 on intestinal B cells. MAIT cells accumulate in the gut (and perhaps the lung) mucosa, where most microorganisms enter the body. Thus, MAIT cells are probably involved in the host response at the site of pathogen entry. IgA secretion in the gut is dependent on microbe-derived signals and does not always require T cells²⁸. Given the parallel absence of IgA secretion and MAIT cells in germ-free mice²⁹, we propose that these invariant T cells may be involved in a negative feedback loop in which microbial products induce both IgA secretion and the cell-surface expression of MR1/ligand. This would activate MAIT cells, which would in turn secrete inhibitory lymphokines, preventing hyperactivation of the gut immune system. Consistent with this, our preliminary results indicate that the proportion of IgA⁺ cells is increased in MR1^{-/-} mice, which lack MAIT cells. Alternatively, MAIT cells may have a non-immune role, as $\gamma\delta$ DETC do by promoting wound healing in the skin³⁰. Regardless, the strict conservation of both MR1 and MAIT cells among mammalian species implies stringent evolutionary pressure for the function(s) fulfilled by MAIT cells. □

Methods

Mice

$\beta 2m$ -deficient (C57Bl/6 N9, B6N9), CD45.1 congenic B6, Ii-deficient (B6/129) and μ MT-deficient (B6 N10) mice were obtained from the Centre National de la Recherche Scientifique animal facility (Centre de Distribution, de Typages et d'Archivage animal (CDTA)). Germ-free mice (C3H) and matched controls were purchased from the CDTA. TAP-deficient mice were obtained from the Jackson laboratory on a B6/129 background. Ii- and TAP-deficient mice were intercrossed to obtain double-deficient mice. γc -deficient mice were obtained from J. Disanto, and crossed to RAG2^{-/-} and $\beta 2m$ ^{-/-} mice to obtain triple-deficient mice (I. Grandjean *et al.*, manuscript in preparation). xid and J β -deficient mice were kindly provided by P. A. Cazenave.

Generation of MR1-knockout mice

Made using standard techniques, the targeting construct was designed to cleanly delete a 1.4-kb fragment containing exons 2 and 3, which encode the $\alpha 1$ and $\alpha 2$ domains of MR1. Mice carrying the MR1 $\alpha 1/\alpha 2$ -domain deletion were intercrossed to generate homozygote knockout mice. All the mice studied had a mixed C57Bl/6 $\times$ 129P2/OlaHsd background (see Supplementary Methods).

Cell preparation

Cell suspensions were prepared from thymus, spleen, PLN and MLN as described³. In some experiments, DN (CD4⁺CD8⁻) LN cells were obtained by negative selection using anti-CD4-fluorescein isothiocyanate (FITC) and anti-CD8-FITC monoclonal antibodies and anti-FITC magnetic beads; after separation on MS columns (Miltenyi Biotec), the effluent was collected as the DN fraction. Gut LPL were prepared as described elsewhere⁶ (see Supplementary Methods).

Cytofluorometry

Four-colour cytometry was performed with directly conjugated antibodies (Pharmingen) according to standard techniques, and analysed on a FACScalibur (Becton Dickinson). The following antibodies were used: anti-CD45.2-FITC, anti-CD3 ϵ -FITC, anti-CD4-phycoerythrin (PE) and -allophycocyanin (APC), anti-CD5-APC, anti-CD8-FITC and -APC, anti-V β 6-PE, anti-CD44-PE, anti-CD45.1-PE, anti-TCR β -cychrome, anti-CD19-PE or -FITC, anti-B220-cychrome, anti-IgD-FITC, anti-CD11c-FITC and anti-IgM-PE (clone R6-60.2). Biotinylated anti-CD45.1 antibodies were revealed with streptavidin-Tricolor or PE (Caltag).

Fetal liver chimaeras

Ten-week-old B6 (CD45.1 or CD45.2) or $\beta 2m$ ^{-/-} mice (CD45.2) were injected intraperitoneally at day -1 and day 0 with 0.3 and 0.4 mg of anti-NK1.1 monoclonal antibody (PK136; kindly provided by B. Rocha). Mice were sublethally irradiated (950 cGy), and injected 6 h later with 5×10^6 fetal liver cells prepared from day-14 B6 (CD45.1 or CD45.2) or $\beta 2m$ ^{-/-} (CD45.2) fetuses. Mice were analysed three months later and chimerism determined using the congenic CD45.1/2 markers. Quantitative PCR (Q-PCR) and polyclonal sequencing were carried out using RNA isolated from LN TCR $\alpha\beta$ ⁺ DN-enriched cells. For each group of mice, the frequency of DN among TCR $\alpha\beta$ ⁺ cells was $\geq 80\%$ after enrichment.

Bone marrow chimaeras

Bone marrow cell suspensions from 12-week-old B6 and $\beta 2m$ ^{-/-} mice were depleted of T cells with anti-Thy-1 antibody and complement. Equal numbers (5×10^6) of $\beta 2m$ ^{-/-} and B6, RAG2^{-/-} or CD3 ϵ ^{-/-} BM cells were mixed and injected intravenously into RAG2^{-/-} γc ^{-/-} $\beta 2m$ ^{-/-} hosts that had been treated two days before with 150 mg kg⁻¹ 5-fluorouracil. Mice were analysed three months later.

Preparation of human samples

Frozen biopsies from the skin, gut or lung of patients with non-immune disorders were cut in 50 μ m slices. RNA was extracted with 1 ml of RNABle (Eurobio) as described³.

Cloning and sequencing of V α 19-J α 33 junctions

CD4⁺CD8⁺ LN T cells were prepared with anti-CD4, -CD8 and -MHC class II antibodies and complement, followed by positive selection with anti-TCR β -FITC (Pharmingen) and anti-FITC magnetic beads (Miltenyi Biotec). V α 19-J α 33 junctions were amplified using nested PCR with the following primers: V α 19 external CACTTTCCTGAGCCGCTCGAA, J α 33 external TATAATTAGCTTGGTCCCAGA, V α 19 internal GCTTCTGACAGAGCT CCAG and J α 33 internal CAGAGCCCCAGATCAACTGAT. Fragments were cloned and sequenced using standard techniques.

Mouse MR1A transfectants

Mouse MR1A cDNA was amplified for 35 cycles from B6 splenocytes using the following primers: AAGAAGGAGATCTGTGATGGTGCTCCTGTACCTCTGCTCG and AGAGAAAGAATTCGAGAGGGAGAGCTTCCCTCATTCACTTG. The PCR product was cloned into the EGFP-N1 vector (Clontech), sequenced and transfected into RBL (rat mastocytoma) or HOM-2 (human B lymphoblastoid) cell lines. After 10–14 days, growing wells were tested for GFP expression and GFP^{hi} cells were FACS-sorted and plated at one cell per well (see Supplementary Methods).

Screen of haematopoietic cell lines and transfectants

T-T hybridomas were made as described³ from DN mesenteric LN T cells from TAP^{-/-}Ii^{-/-} mice. Cell lines or stable transfectants to be screened were irradiated and distributed at 5×10^4 cells per well in flat-bottomed 96-well plates and 5×10^4 hybridoma were added. After a 48-h incubation, the supernatant was harvested and tested for IL-2 content using the IL-2-dependent CTLL2 cell line. For blocking experiments, serial dilutions of anti-TCR β monoclonal antibody (anti-H57 NA/LE, Pharmingen) or hamster isotype control (clone Ha4/8) were added to the wells (see Supplementary Methods).

Processing of human blood samples, molecular biology methods, RNA extraction, reverse transcription, TCR primers and PCR methods have been described previously³. For clarity, all Q-PCR data have been normalized to C α expression by shifting the V α -J α PCR curves along the x axis.

Received 26 September 2002; accepted 10 January 2003; doi:10.1038/nature01433.

- Bendelac, A., Rivera, M. N., Park, S. H. & Roark, J. H. Mouse CD1-specific NK1 T cells: development, specificity, and function. *Annu. Rev. Immunol.* **15**, 535–562 (1997).
- Kronenberg, M. & Gapin, L. The unconventional lifestyle of NKT cells. *Nature Rev. Immunol.* **2**, 557–568 (2002).
- Tilloy, F. *et al.* An invariant T cell receptor α chain defines a novel TAP-independent major histocompatibility complex class Ib-restricted α/β T cell subpopulation in mammals. *J. Exp. Med.* **189**, 1907–1921 (1999).
- Riegert, P., Wanner, V. & Bahram, S. Genomics, isoforms, expression, and phylogeny of the MHC class I-related MR1 gene. *J. Immunol.* **161**, 4066–4077 (1998).
- Berlin, C. *et al.* $\alpha 4\beta 7$ integrin mediates lymphocyte binding to the mucosal vascular addressin MAdCAM-1. *Cell* **74**, 185 (1993).
- Arstila, T. *et al.* Identical T cell clones are located within the mouse gut epithelium and lamina propria and circulate in the thoracic duct lymph. *J. Exp. Med.* **191**, 823–834 (2000).
- Tourne, S. *et al.* Biosynthesis of major histocompatibility complex molecules and generation of T cells in Ii TAP1 double-mutant mice. *Proc. Natl Acad. Sci. USA* **93**, 1464–1469 (1996).
- Bix, M. & Raulet, D. Inefficient positive selection of T cells directed by haematopoietic cells. *Nature* **359**, 330–333 (1992).
- Bendelac, A. Positive selection of mouse NK1⁺ T cells by CD1-expressing cortical thymocytes. *J. Exp. Med.* **182**, 2091–2096 (1995).
- Urdahl, K. B., Sun, J. C. & Bevan, M. J. Positive selection of MHC class Ib-restricted CD8⁺ T cells on haematopoietic cells. *Nature Immunol.* **3**, 772–779 (2002).
- Jakovcic, A. *et al.* Analysis of homozygous mutant chimeric mice: deletion of the immunoglobulin heavy-chain joining region blocks B-cell development and antibody production. *Proc. Natl Acad. Sci. USA* **90**, 2551–2555 (1993).
- Kitamura, D., Roes, J., Kuhn, R. & Rajewsky, K. A B cell-deficient mouse by targeted disruption of the membrane exon of the immunoglobulin μ chain gene. *Nature* **350**, 423–426 (1991).
- Macpherson, A. J. *et al.* IgA production without μ or δ chain expression in developing B cells. *Nature Immunol.* **2**, 625–631 (2001).
- Berland, R. & Wortis, H. H. Origins and functions of B-1 cells with notes on the role of CD5. *Annu. Rev. Immunol.* **20**, 253–300 (2002).
- Tsukada, S., Rawlings, D. J. & Witte, O. N. Role of Bruton's tyrosine kinase in immunodeficiency. *Curr. Opin. Immunol.* **6**, 623–630 (1994).
- Maenaka, K. & Jones, E. Y. MHC superfamily structure and the immune system. *Curr. Opin. Struct. Biol.* **9**, 745–753 (1999).
- Ugolini, S. & Vivier, E. Multifaceted roles of MHC class I and MHC class I-like molecules in T cell activation. *Nature Immunol.* **2**, 198–200 (2001).
- Bahram, S. MIC genes: from genetics to biology. *Adv. Immunol.* **76**, 1–60 (2000).
- Hashimoto, K., Hirai, M. & Kurosawa, Y. A gene outside the human MHC related to classical HLA class I genes. *Science* **269**, 693–695 (1995).
- Gu, H., Forster, I. & Rajewsky, K. Sequence homologies, N sequence insertion and JH gene utilization in VHDJH joining: implications for the joining mechanism and the ontogenetic timing of Ly1 B cell and B-CLL progenitor generation. *EMBO J.* **9**, 2133–2140 (1990).
- Itoharu, S. *et al.* T cell receptor δ gene mutant mice: independent generation of $\alpha\beta$ T cells and programmed rearrangements of $\gamma\delta$ TCR genes. *Cell* **72**, 337–348 (1993).
- Asarnow, D. M., Cado, D. & Raulet, D. H. Selection is not required to produce invariant T-cell receptor γ -gene junctional sequences. *Nature* **362**, 158–160 (1993).
- Mallick-Wood, C. A. *et al.* Conservation of T cell receptor conformation in epidermal $\gamma\delta$ cells with disrupted primary V γ gene usage. *Science* **279**, 1729–1733 (1998).
- Park, S. H., Benlagha, K., Lee, D., Balish, E. & Bendelac, A. Unaltered phenotype, tissue distribution and function of V α 14⁺ NKT cells in germ-free mice. *Eur. J. Immunol.* **30**, 620–625 (2000).

25. Groh, V. *et al.* Cell stress-regulated human major histocompatibility complex class I gene expressed in gastrointestinal epithelium. *Proc. Natl Acad. Sci. USA* **93**, 12445–12450 (1996).
26. Wang, C. R. *et al.* Nonclassical binding of formylated peptide in crystal structure of the MHC class Ib molecule H2-M3. *Cell* **82**, 655–664 (1995).
27. Lee, N., Goodlett, D. R., Ishitani, A., Marquardt, H. & Geraghty, D. E. HLA-E surface expression depends on binding of TAP-dependent peptides derived from certain HLA class I signal sequences. *J. Immunol.* **160**, 4951–4960 (1998).
28. Macpherson, A. J. *et al.* A primitive T cell-independent mechanism of intestinal mucosal IgA responses to commensal bacteria. *Science* **288**, 2222–2226 (2000).
29. Benveniste, J., Lespinats, G. & Salomon, J. Serum and secretory IgA in axenic and holoxenic mice. *J. Immunol.* **107**, 1656–1662 (1971).
30. Jameson, J. *et al.* A role for skin $\gamma\delta$ cells in wound repair. *Science* **296**, 747–749 (2002).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements We thank S. Laigneau, N. Froux, M. Garcia, F. Valette and I. Cissé for managing the mouse colonies in Paris, E. Wagner and colleagues for animal care at the Basel Institute for Immunology, C. DeSouza for help in the membrane biotinylation, Z. Maciorowski for cell sorting, and S. Kuschert and A. Dierich for blastocyst injection. We thank F. Ledest for patient blood samples, N. Brousse and F. Geissman for human biopsies, P. A. Cazenave's group for xid and J β knockout mice, and K. Rajewsky for the Cre transgenic mice. We thank M. Bonneville, S. Amigorena, C. Thery, P. Benaroch, M. Colonna, D. Freemont and T. Hansen for discussions and for reviewing the manuscript. This work was supported by grants from the Association de la Recherche Contre le Cancer, Fondation pour la Recherche Médicale, Institut National de la Santé et de la Recherche Médicale and Section Médicale de l'Institut Curie. S.G. thanks Hoffmann la Roche for supporting the Basel Institute for Immunology and M. Colonna for support in St Louis.

Authors' contributions S. Gilfillan and O. Lantz share senior authorship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to O.L. (e-mail: olivier.lantz@curie.net) or S.G. (e-mail: susang@pathbox.wustl.edu).

Loss of integrin $\alpha\beta 6$ -mediated TGF- β activation causes Mmp12-dependent emphysema

David G. Morris^{*,†}, Xiaozhu Huang^{*}, Naftali Kaminski[‡], Yanli Wang^{*}, Steven D. Shapiro[§], Gregory Dolganov[†], Adam Glick^{||} & Dean Sheppard^{*,†}

^{*} Lung Biology Center, Department of Medicine, San Francisco General Hospital, and [†] Division of Pulmonary and Critical Care Medicine, Department of Medicine, University of California, San Francisco, California 94143, USA

[‡] Department of Medicine, Division of Pulmonary and Critical Care Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA

[§] Department of Medicine, Section of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA

^{||} Laboratory of Cellular Carcinogenesis and Tumor Promotion, National Cancer Institute, Bethesda, Maryland 20892, USA

Integrins are heterodimeric cell-surface proteins that regulate cell growth, migration and survival. We have shown previously that the epithelial-restricted integrin $\alpha\beta 6$ has another critical function; that is, it binds and activates latent transforming growth factor- β (TGF- β)^{1,2}. Through a global analysis of pulmonary gene expression in the lungs of mice lacking this integrin (*Itgb6* null mice) we have identified a marked induction of macrophage metalloelastase (Mmp12)—a metalloproteinase that preferentially degrades elastin and has been implicated in the chronic lung disease emphysema³. Here we report that *Itgb6*-null mice develop age-related emphysema that is completely abrogated either by transgenic expression of versions of the $\beta 6$ integrin subunit that support TGF- β activation, or by the loss of Mmp12. Furthermore, we show that the effects of *Itgb6* deletion are overcome by simultaneous transgenic expression of

active TGF- $\beta 1$. We have uncovered a pathway in which the loss of integrin-mediated activation of latent TGF- β causes age-dependent pulmonary emphysema through alterations of macrophage Mmp12 expression. Furthermore, we show that a functional alteration in the TGF- β activation pathway affects susceptibility to this disease.

Pulmonary emphysema, which is characterized by simplification of alveolar architecture, loss of lung elasticity, and enlargement of alveolar airspaces, is a worldwide health problem largely attributable to exposure to tobacco smoke. Extracellular proteases, which regulate extracellular matrix homeostasis, have been implicated in tobacco-smoke-induced pulmonary emphysema. Nevertheless, the regulation of these proteases *in vivo* is poorly understood and a regulatory role of lung epithelial cells in this process has yet to be described.

Mice that do not express the $\beta 6$ subunit of the $\alpha\beta 6$ integrin (*Itgb6*[−]) spontaneously develop increased expression of the extracellular macrophage metalloproteinase Mmp12 (macrophage metalloelastase) in their lungs by eight weeks of age. In fact, on the basis of whole-organ gene expression profiling using the

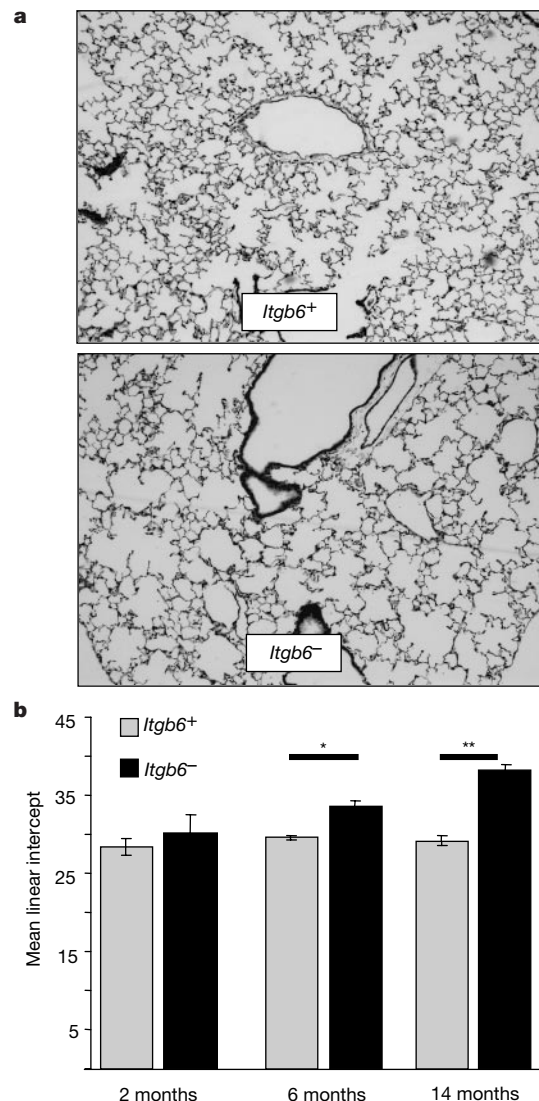


Figure 1 Spontaneous, progressive pulmonary emphysema in *Itgb6*[−] mice.

a, Representative histological sections ($\times 10$ objective) of lungs from *Itgb6*^{+/+} and *Itgb6*^{−/−} mice at 14 months of age show enlarged alveoli indicative of emphysema in *Itgb6*^{−/−} mice.

b, Mean linear intercepts ($\mu\text{m} \pm \text{s.e.m.}$) of alveolar septae measured in the lungs of five *Itgb6*^{+/+} (wild type) and five *Itgb6*^{−/−} mice at 2, 6 and 14 months of age. Asterisk, $P = 0.0006$; double asterisk, $P = 0.0002$.

Affymetrix mu6500 microarray, expression of *Mmp12* is increased in the lungs of *Itgb6*⁻ mice to at least 18-fold above the level of expression in wild-type mice³ (Supplementary Fig. 1). Using real-time quantitative polymerase chain reaction (PCR) analysis, we found that alveolar macrophages from the lungs of *Itgb6*⁻ mice expressed 200-fold more *Mmp12* messenger RNA than alveolar macrophages from wild-type mice (ratio of *Mmp12* mRNA

abundance in *Itgb6*⁻ compared with *Itgb6*⁺ mice = 228 ± 20; *P* = 0.0008). Casein zymography confirmed increased levels of *Mmp12* protein in the bronchoalveolar lavage fluid of *Itgb6*⁻ mice by eight weeks of age (data not shown).

Mmp12 is an extracellular matrix-degrading metalloproteinase that is expressed only by tissue macrophages and placental trophoblasts⁴⁻⁶. Mice lacking *Mmp12* (*Mmp12*^{-/-}) do not develop alveolar enlargement—the definitive characteristic of pulmonary emphysema—after exposure to cigarette smoke, suggesting that *Mmp12* is important in the development of this disease⁷. Our observation that *Mmp12* gene expression is markedly increased in the lungs of *Itgb6*⁻ mice also raised the possibility that loss of the αvβ6 integrin could result in the development of emphysema over time. To test this hypothesis, we examined alveolar size in lungs of *Itgb6*⁻ and wild-type mice by measuring the mean linear intercept of alveolar septae at 2, 6 and 14 months of age. We found that *Itgb6*⁻ mice, despite having normal alveolar size at 2 months of age, spontaneously developed progressive alveolar enlargement, or pulmonary emphysema, over time, whereas wild-type mice did not (Fig. 1).

The αvβ6 integrin has two distinct functions *in vivo* that depend on specific regions of the β6 subunit: enhancement of epithelial cell proliferation, which depends on the carboxy-terminal 11 amino acids⁸, and the activation of latent TGF-β, which is independent of these residues¹. Emphysema in *Itgb6*-deleted mice might conceivably be due to a loss of either of these functions. To determine which of the functions of the β6 integrin subunit is critical to prevent the development of spontaneous emphysema, we studied alveolar size, *Mmp12* expression levels and alveolar macrophage morphology in

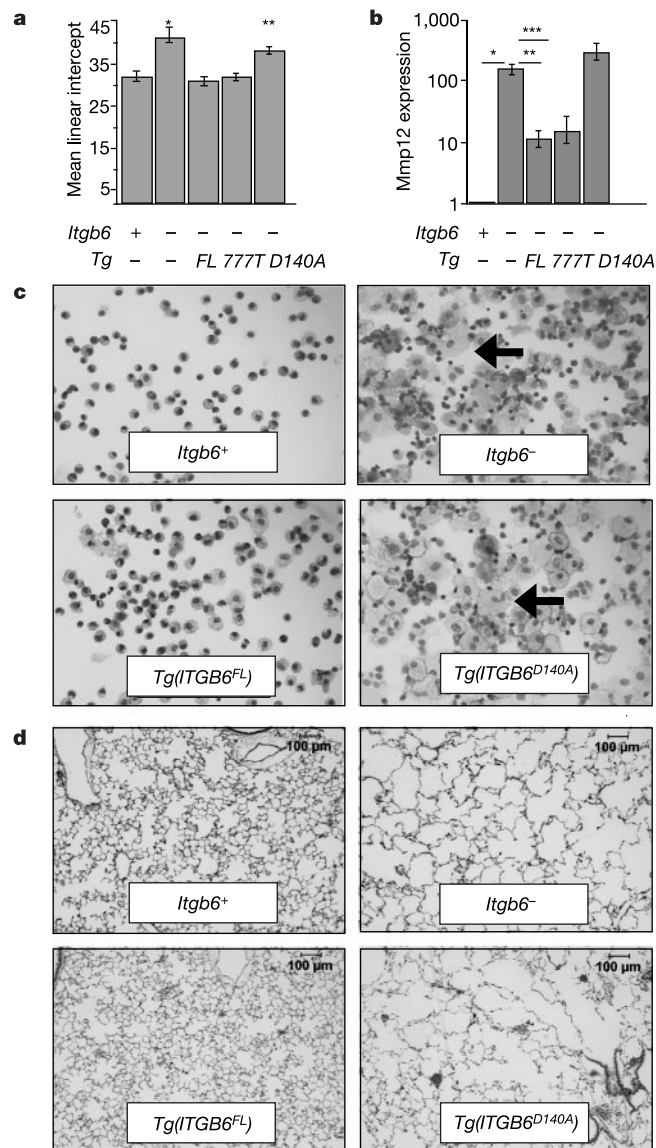


Figure 2 Transgenic expression of human *ITGB6*^{FL} or *ITGB6*^{777T} reduces *Mmp12* expression, normalizes macrophage appearance and prevents the development of airspace enlargement in *Itgb6*⁻ mice. **a**, Mean linear intercepts in 9-month-old mice show that *Itgb6*⁻ and *Itgb6*⁻ *Tg*(*ITGB6*^{D140A}) mice develop alveolar enlargement whereas *Itgb6*⁻ *Tg*(*ITGB6*^{FL}) and *Itgb6*⁻ *Tg*(*ITGB6*^{777T}) mice do not. Asterisk, *P* = 0.002; double asterisk, *P* = 0.004 compared with wild-type mice; *n* = 5 per group. **b**, Two-month-old *Itgb6*⁺, *Itgb6*⁻ *Tg*(*ITGB6*^{FL}) and *Itgb6*⁻ *Tg*(*ITGB6*^{777T}) mice have markedly reduced levels of *Mmp12* expression compared with *Itgb6*⁻ and *Itgb6*⁻ *Tg*(*ITGB6*^{D140A}) mice. Asterisk, *P* = 2.3 × 10⁻⁷; double asterisk, *P* = 0.0001; triple asterisk, *P* = 0.0002; *n* = 4 per group. **c**, Representative cytopsin concentrates from bronchoalveolar lavage fluid (× 40 objective) from mice of each genotype show abundant, large vacuolated alveolar macrophages from both *Itgb6*⁻ and *Itgb6*⁻ *Tg*(*ITGB6*^{D140A}) mice (arrows). Alveolar macrophages from *Itgb6*⁻ *Tg*(*ITGB6*^{777T}) mice are nearly normal and indistinguishable from those recovered from *Itgb6*⁻ *Tg*(*ITGB6*^{FL}) mice (data not shown). **d**, Representative histological sections (× 10 objective) of the lungs of *Itgb6*⁺, *Itgb6*⁻, *Itgb6*⁻ *Tg*(*ITGB6*^{FL}) and *Itgb6*⁻ *Tg*(*ITGB6*^{D140A}) mice.

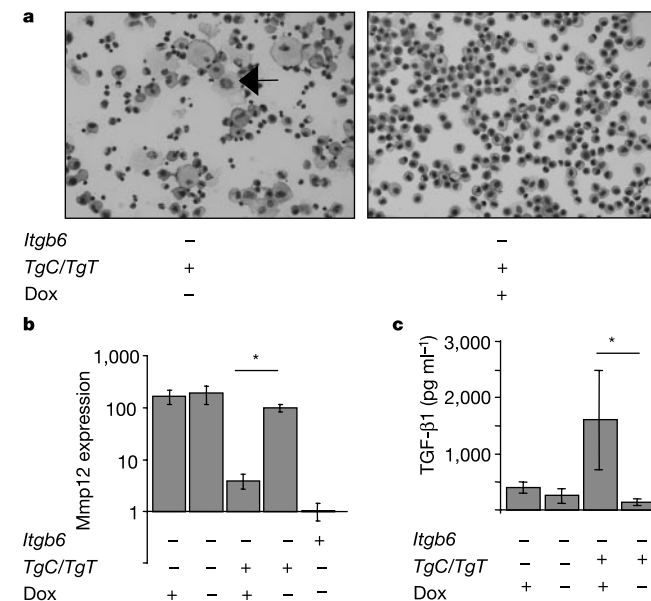


Figure 3 Transgenic expression of *Tgfb1Cys-Ser*^{223,225} normalizes alveolar macrophage appearance and reduces *Mmp12* expression in *Itgb6*⁻ mice. **a**, Representative cytopsin concentrates of bronchoalveolar lavage fluid (× 40 objective) from *Itgb6*⁻ double-transgenic mice show abundant, large vacuolated alveolar macrophages (arrow) from mice without transgene induction (Dox -) and normal-appearing macrophages from mice with transgene induction (Dox +). *TgC/TgT*, *Tg*(*CCSP-rTA*) *Tg*(*tetO-Tgfb1Cys-Ser*^{223,225}). Alveolar macrophages from water- and doxycycline-treated *Itgb6*⁻ *TgC*⁻ *TgT*⁻ mice and water-treated *Itgb6*⁻ *TgC*⁺ *TgT*⁺ mice are identical in appearance (data not shown). **b**, Measurement of *Mmp12* transcript abundance in bronchoalveolar lavage fluid cells from 6-week-old mice shows that *Mmp12* expression in *Itgb6*⁻ *TgC*⁺ *TgT*⁺ mice is reduced nearly 45-fold by 21 days of doxycycline-induced gene expression. Asterisk, *P* = 4 × 10⁻⁶; *n* = 3 per group. **c**, Total TGF-β1 in bronchoalveolar lavage fluid in *Itgb6*⁻ *TgC*⁺ *TgT*⁺ mice is increased after 14 days of doxycycline treatment. Asterisk, *P* = 0.03; *n* = 3–5 per group.

Itgb6[−] mice that expressed transgenic versions of either full-length human β6 (*Itgb6*[−] *Tg(ITGB6^{FL})*), a truncated form of the human β6 integrin subunit lacking the C-terminal 11 amino acids (*Itgb6*[−] *Tg(ITGB6^{777T})*), or a ligand-binding-incompetent form of β6 (*Itgb6*[−] *Tg(ITGB6^{D140A})*) in distal bronchiolar and type II lung epithelial cells^{9,10}. We confirmed lung-specific expression of these proteins *in vivo* by western blotting (data not shown) and have shown previously that each of these mutants is well expressed on the surface of epithelial cells *in vitro*^{9,10}. Transgenic introduction of the full-length human β6 integrin subunit (*Tg(ITGB6^{FL})*) into *Itgb6*[−] mice eliminated the development of spontaneous emphysema, substantially reduced the induction of Mmp12, and prevented the development of vacuolated alveolar macrophages (Fig. 2). Transgenic introduction of the β6 truncation mutant (*Tg(ITGB6^{777T})*)—a mutant that cannot support enhanced cell growth in three-dimensional culture or *in vivo*⁸ but that can bind the latency-associated protein (LAP) and activate TGF-β¹—also eliminated development of spontaneous emphysema, substantially reduced the induction of Mmp12, and prevented the development of vacuolated alveolar macrophages in *Itgb6*[−] mice. However, transgenic introduction of a ligand-binding-incompetent mutant of the β6 protein, which can neither bind nor activate TGF-β as a result of an aspartic acid to alanine mutation at position 140 (D140A) in the metal-ion-dependent adhesion site (MIDAS) domain of β6 (*Tg(ITGB6^{D140A})*), did not prevent the development of spontaneous emphysema, the induction of Mmp12 expression, or the development of vacuolated alveolar macrophages in *Itgb6*[−] mice.

These results indicate that although expression of the β6 integrin subunit by epithelial cells in the lung is sufficient to prevent the development of spontaneous emphysema in ageing mice, the prevention of spontaneous emphysema in *Itgb6*-deleted mice seems to be dependent on the ability of the transgenic β6 integrin to bind and activate latent TGF-β. These findings further suggested that the alveolar macrophage phenotype in *Itgb6*-deleted mice was due, in part or in whole to a chronic deficiency of active TGF-β which, in turn, was caused by a defect in the activation of latent TGF-β by the αvβ6 integrin. This possibility is further strengthened

by *in vitro* studies showing Smad-3-dependent regulation of Mmp12 expression by TGF-β in peripheral blood-derived macrophages^{11,12}.

Therefore, we investigated whether expression of active TGF-β itself in the lungs of *Itgb6*-deleted mice was sufficient to bypass the activation defect imposed by *Itgb6* deficiency and thereby prevent the development of the early hallmarks of the *Itgb6*-deleted pulmonary phenotype (that is, increased Mmp12 expression and the accumulation of large vacuolated alveolar macrophages). To do this, we first generated mice that expressed constitutively active TGF-β1 under positive tetracycline-regulated transcriptional control (*Itgb6*[−] *Tg(CCSP-rtTA)* *Tg(tetO-Tgfb1Cys-Ser^{223,225})*). We found that these mice, when bred to be double allelic for both transgenes, had detectable levels of TGF-β1 in bronchoalveolar lavage fluid after two weeks of doxycycline administration, whereas transgene-negative mice, or isogenic mice receiving water, did not (data not shown). We then bred copies of both of these transgenes into syngeneic *Itgb6*[−] mice. We found that double-transgenic *Itgb6*[−] mice, when given doxycycline to cause expression of constitutively active TGF-β1 in their lungs, had alveolar macrophages that appeared normal and had substantially reduced Mmp12 expression compared with untreated, double-transgenic *Itgb6*[−] mice (Fig. 3). Furthermore, expression of active TGF-β1 eliminated the lymphocyte and neutrophil accumulation in the alveoli of *Itgb6*[−] mice (data not shown). These findings show that expression of active TGF-β1 in the lungs of *Itgb6*[−] mice corrects the principal abnormalities caused by deletion of the integrin gene—presumably by bypassing the associated defect in latent TGF-β activation.

To determine whether the induction of Mmp12 expression in *Itgb6*[−] mice is responsible for the alveolar enlargement that we observed, we generated mice deficient in the expression of both β6 and Mmp12 (*Itgb6*[−] *Mmp12*[−]). By 8 months of age, *Itgb6*[−] mice had developed significant alveolar enlargement whereas *Itgb6*[−] *Mmp12*[−] mice displayed no alveolar enlargement (Fig. 4a, c). These results show that Mmp12 is essential for the development of emphysema in *Itgb6*[−] mice. Notably, the prevention of spontaneous emphysema in *Itgb6*[−] *Mmp12*[−] mice was not due to a

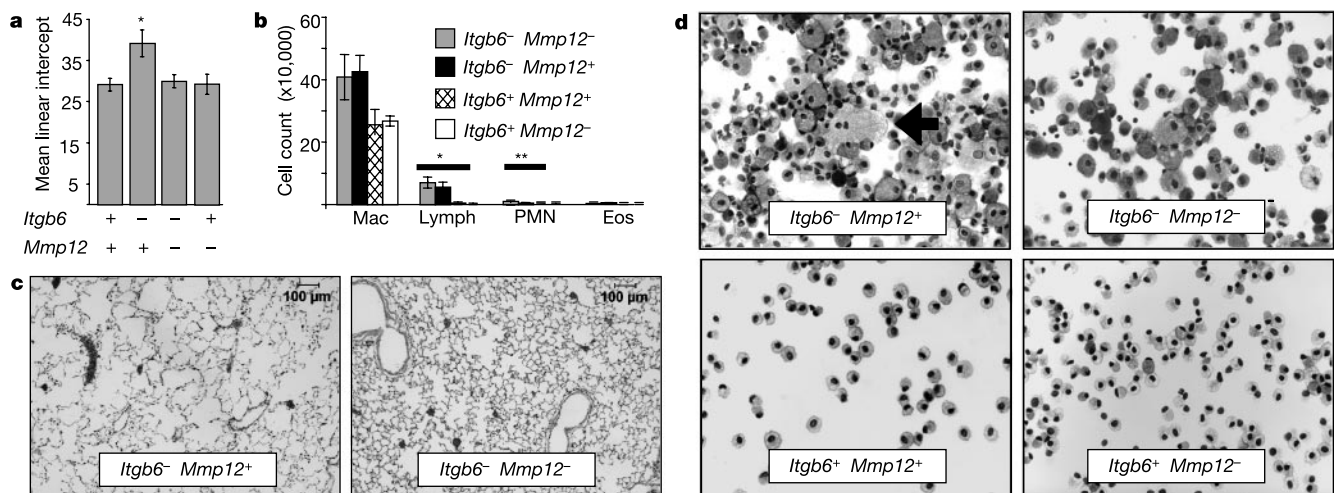


Figure 4 Deletion of Mmp12 prevents spontaneous emphysema in *Itgb6*[−] mice but does not affect inflammation. **a**, Mean linear intercepts in the lungs of 8-month-old *Itgb6*⁺ *Mmp12*⁺, *Itgb6*[−] *Mmp12*⁺, *Itgb6*[−] *Mmp12*[−] and *Itgb6*⁺ *Mmp12*[−] mice show that alveolar enlargement in *Itgb6*[−] mice is prevented by simultaneous deletion of Mmp12. Asterisk, $P = 0.009$ for *Itgb6*[−] *Mmp12*⁺ compared with *Itgb6*[−] *Mmp12*[−] mice; $n = 5$ per group. **b**, Comparison of total cell counts in bronchoalveolar lavage fluid from 3-month-old *Itgb6*[−] *Mmp12*⁺, *Itgb6*[−] *Mmp12*[−], *Itgb6*⁺ *Mmp12*⁺ and *Itgb6*⁺ *Mmp12*[−] mice reveals similar degrees of macrophage, polymorphonuclear leukocyte and lymphocyte accumulation in *Itgb6*[−] *Mmp12*⁺ and *Itgb6*[−] *Mmp12*[−] mice. Mac,

alveolar macrophages; lymph, lymphocytes; PMN, polymorphonuclear leukocytes; eos, eosinophils. Asterisk, $P < 0.001$; double asterisk, $P = 0.02$ for *Itgb6*[−] *Mmp12*[−] compared with wild-type mice; $n = 5$ –8 per group. **c**, Representative histological sections ($\times 10$ objective) of the lungs of 8-month-old mice show airspace enlargement in *Itgb6*[−] *Mmp12*⁺ but not *Itgb6*[−] *Mmp12*[−] mice. **d**, Representative cytospin concentrates from bronchoalveolar lavage fluid ($\times 40$ objective) from mice of each genotype show abundant, large vacuolated alveolar macrophages in both *Itgb6*[−] *Mmp12*⁺ and *Itgb6*[−] *Mmp12*[−] mice (arrow).

reduction in the degree of lung inflammation, because, by means of bronchoalveolar lavage of the lungs of these double knockout mice, we found that the degree of lung inflammation was similar to that seen in *Itgb6*^{-/-} mice (Fig. 4b)^{10,13}. In addition, *Itgb6*^{-/-} *Mmp12*^{-/-} mice still developed patchy juvenile baldness and the associated skin inflammation that we have previously reported in *Itgb6*^{-/-} mice (data not shown)¹³. Therefore, *Mmp12* activity is not required for inflammatory cell recruitment into the lungs or skin of *Itgb6*^{-/-} mice, and lung inflammation, in the absence of *Mmp12*, is not sufficient to cause emphysema in *Itgb6*^{-/-} mice.

Several recent reports have described the development of emphysema in rodent models. These include mice transgenic for interferon- γ (IFN- γ)¹⁴, tumour necrosis factor- α (TNF- α)¹⁵ or interleukin-13 (IL-13)¹⁶, mice deficient in surfactant protein D¹⁷ or tissue inhibitor of metalloproteinase 3 (Timp-3)¹⁸, and rats treated with an inhibitor of vascular endothelial-derived growth factor receptor 2 (Vegfr-2)¹⁹. In the mouse models involving IL-13, IFN- γ , TNF- α and surfactant protein D, rapidly progressive emphysema follows induction of a multitude of elastolytic and collagenolytic enzymes. Although our results do not eliminate the possibility that other proteases may interact with *Mmp12* in the development of emphysema, the absence of a significant induction of the expression of other proteases in the lungs as measured by microarray analysis, and the lack of emphysema in *Itgb6*^{-/-} *Mmp12*^{-/-} mice indicates that *Mmp12* is central to this process in *Itgb6*^{-/-} mice.

We demonstrate a new *in vivo* pathway in which the loss of an epithelial integrin, which is known to cause a local deficiency in active TGF- β ^{1,2}, results in increased expression of *Mmp12* by alveolar macrophages and causes emphysema over time. We also show that the increased expression of *Mmp12* can be prevented by expression of constitutively active TGF- β in these integrin-deficient mice. The gradual, age-related emphysema in *Itgb6*^{-/-} mice, in contrast to the severe and rapidly progressive forms observed in mice that overexpress IL-13, TNF- α and IFN- γ , is more akin to the pace of emphysema commonly observed in humans. Furthermore, the pathological features of *Itgb6*^{-/-} mice resemble those previously reported in young tobacco smokers²⁰. Although cigarette smoke is the major environmental factor responsible for causing emphysema in humans, only a minority of heavy smokers develop emphysema, which suggests that other risk factors are important. Family studies of patients with emphysema have demonstrated clearly the importance of genetic factors in determining an individual's susceptibility to this disease. Although congenital deficiency of α_1 -antitrypsin has been shown to increase susceptibility to emphysema, it is a rare condition, and most of the genetic factors affecting susceptibility remain unexplained²¹. The finding that mice lacking the $\alpha\beta 6$ integrin had persistently elevated expression of *Mmp12* and developed slowly progressive, age-related emphysema, suggests that the integrity of this pathway may also be important in preventing emphysema in humans. Our results suggest that abnormalities in any of the steps in this pathway of TGF- β activation or signalling (for example, decreased $\alpha\beta 6$ expression or function, decreased expression or function of TGF- β , or defects in the TGF- β signalling pathway in macrophages) may contribute to genetic or acquired susceptibility to this common and debilitating disease. □

Methods

Itgb6^{-/-}, *Mmp12*^{-/-} and transgenic mice

Itgb6^{-/-} mice were generated as described on a 129Sv genetic background¹³. *Mmp12*^{-/-} mice were generated as described on a 129Sv genetic background⁷. Transgenic mice expressing a truncated form of the human $\beta 6$ subunit lacking the last 11 amino acid residues of the cytoplasmic domain (*Tg(ITGB6*^{777T}) or an aspartic acid to alanine point mutant (*Tg(ITGB6*^{D140A})) under transcriptional control of the tissue-specific surfactant protein C (Spc) promoter were generated using methods as described¹⁰. The 777T fragment was excised from pCDNA1Neo with *XhoI* and *XbaI*, blunt ends were created, and it was cloned into the expression plasmid pUC18SPC3.7 (a gift of J. Whitsett), which had been digested with *BglII* and blunt-ended^{22,23}. The D140A fragment was subcloned into pMamBlue, digested with *XhoI* and *XbaI*, excised with *SalI* and cloned into

pUC18SPC3.7 that had been digested with *SalI*. Transgene incorporation in founder lines was confirmed by Southern blot analysis, and tissue-specific protein expression was confirmed using western analysis using the 4B5 rabbit monoclonal anti-human $\beta 6$ antibody¹⁰. These transgenic mice were backcrossed five generations onto an *Itgb6*^{-/-} C57BL/6 background. *Tg(CCSF-rtTA)Tg(tetO-Tgfb1Cys-Ser*^{223,225}) double-transgenic mice were generated by breeding together two independent lines of FVB mice carrying either the *Tg(CCSF-rtTA)* transgenic construct (a gift of J. Whitsett)²⁴ or the *Tg(tetO-Tgfb1Cys-Ser*^{223,225}) transgenic construct²⁵. Mice carrying copies of both transgenes were verified by PCR genotyping and bred onto an FVB *Itgb6*^{-/-} genetic background. Doxycycline-inducible expression of TGF- β 1 was verified by enzyme-linked immunosorbent assay (Pharmingen).

Quantitative real-time RT-PCR

RNA isolation, treatment, primer design and amplicon detection probe design were carried out as described²⁶. All primers and probe sequences are available at <http://asthmagenomics.ucsf.edu/pubs/>. The mean number of cycles to threshold (C_T) of fluorescence detection was calculated for each sample and the results were normalized to the mean C_T of glyceraldehyde 3-phosphate dehydrogenase (Gapdh) for each sample tested. Results are expressed as a fold increase (or decrease) in complementary DNA abundance compared with control animals. We used the normalized C_T differences ($Mmp12\ C_T - Gapdh\ C_T$) for all statistical comparisons.

Quantitative morphometry and mean linear intercept

The trachea and lungs were removed together and inflated with 10% buffered neutral formalin (VWR Scientific) to 25 cm water pressure. Parasagittally sectioned tissue was embedded in paraffin, and 5- μ m-thick sections were stained with haematoxylin and eosin. Slide images were digitally captured using CAST-grid software version 2.00.1 (Olympus) with a line-counting tool at $\times 20$ magnification with meander sampling beginning at a randomly selected point, sampling all of the tissue in an unbiased fashion. The mean linear intercept (defined as the linear sum of the lengths, in μ m, of all lines in all frames counted divided by the number of intercepts (defined as an alveolar septa intersecting with a counting line)) was calculated according to an adaptation of the method of ref. 27. A minimum of 12 fields, 200 intercepts and 300 points for each animal were measured, sampling all lobes.

Bronchoalveolar lavage and cytospin analysis

Animals were terminally anaesthetized with Methoxyfluorane (Metofane), their tracheae cannulated, and lungs washed with five sequential aliquots of 0.8 ml PBS at room temperature. Aliquots were pooled, centrifuged, and the cell pellets re-suspended in red-blood-cell lysis buffer (Sigma). Cells were counted using a haemocytometer and cell concentrates were stained with Diff-Quick (Dade Diagnostics). Cell subsets were counted under $\times 40$ objective magnification (300 cells per slide).

Statistical analysis

All data are reported as mean $\pm$ s.e.m. All between-group comparisons of mean linear intercepts were made using the Kruskal–Wallis test, and subsequently the Mann–Whitney *U*-test, using the Bonferroni correction for multiple comparisons. All between-group comparisons of normalized C_T differences, TGF- β protein levels and cell counts were made using analysis of variance followed by Bonferroni-corrected unpaired *t*-tests. Only experimentally relevant comparisons are reported. Comparisons with an α -value of $P < 0.05$ were assigned as significantly different. Statistical analyses were conducted using SYSTAT version 9.0 software (SPSS Science).

Received 6 September; accepted 23 December 2002; doi:10.1038/nature01413.

- Munger, J. S. *et al.* The integrin $\alpha\beta 6$ binds and activates latent TGF β 1: a mechanism for regulating pulmonary inflammation and fibrosis. *Cell* **96**, 319–328 (1999).
- Pittet, J. F. *et al.* TGF- β is a critical mediator of acute lung injury. *J. Clin. Invest.* **107**, 1537–1544 (2001).
- Kaminski, N. *et al.* Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis. *Proc. Natl Acad. Sci. USA* **97**, 1778–1783 (2000).
- Werb, Z. & Gordon, S. Elastase secretion by stimulated macrophages. Characterization and regulation. *J. Exp. Med.* **142**, 361–377 (1975).
- Shapiro, S. D. *et al.* Molecular cloning, chromosomal localization, and bacterial expression of a murine macrophage metalloelastase. *J. Biol. Chem.* **267**, 4664–4671 (1992).
- Belaouaj, A. *et al.* Human macrophage metalloelastase. Genomic organization, chromosomal location, gene linkage, and tissue-specific expression. *J. Biol. Chem.* **270**, 14568–14575 (1995).
- Hautamaki, R. D., Kobayashi, D. K., Senior, R. M. & Shapiro, S. D. Requirement for macrophage elastase for cigarette smoke-induced emphysema in mice. *Science* **277**, 2002–2004 (1997).
- Agrez, M., Chen, A., Cone, R. L., Pytela, R. & Sheppard, D. The $\alpha\beta 6$ integrin promotes proliferation of colon carcinoma cells through a unique region of the beta 6 cytoplasmic domain. *J. Cell Biol.* **127**, 547–556 (1994).
- Huang, X. Z., Chen, A., Agrez, M. & Sheppard, D. A point mutation in the integrin $\beta 6$ subunit abolishes both $\alpha\beta 6$ binding to fibronectin and receptor localization to focal contacts. *Am. J. Respir. Cell. Mol. Biol.* **13**, 245–251 (1995).
- Huang, X., Wu, J., Zhu, W., Pytela, R. & Sheppard, D. Expression of the human integrin $\beta 6$ subunit in alveolar type II cells and bronchiolar epithelial cells reverses lung inflammation in $\beta 6$ knockout mice. *Am. J. Respir. Cell. Mol. Biol.* **19**, 636–642 (1998).
- Feinberg, M. W. *et al.* Transforming growth factor- β 1 inhibits cytokine-mediated induction of human metalloelastase in macrophages. *J. Biol. Chem.* **275**, 25766–25773 (2000).
- Werner, E. *et al.* Transforming growth factor- β 1 inhibition of macrophage activation is mediated via Smad3. *J. Biol. Chem.* **275**, 36653–36658 (2000).
- Huang, X. Z. *et al.* Inactivation of the integrin $\beta 6$ subunit gene reveals a role of epithelial integrins in regulating inflammation in the lung and skin. *J. Cell Biol.* **133**, 921–928 (1996).

14. Wang, Z. *et al.* Interferon gamma induction of pulmonary emphysema in the adult murine lung. *J. Exp. Med.* **192**, 1587–1600 (2000).
15. Fujita, M. *et al.* Overexpression of tumour necrosis factor- α produces an increase in lung volumes and pulmonary hypertension. *Am. J. Physiol. Lung Cell Mol. Physiol.* **280**, L39–L49 (2001).
16. Zheng, T. *et al.* Inducible targeting of IL-13 to the adult lung causes matrix metalloproteinase- and cathepsin-dependent emphysema. *J. Clin. Invest.* **106**, 1081–1093 (2000).
17. Wert, S. E. *et al.* Increased metalloproteinase activity, oxidant production, and emphysema in surfactant protein D gene-inactivated mice. *Proc. Natl Acad. Sci. USA* **97**, 5972–5977 (2000).
18. Leco, K. J. *et al.* Spontaneous air space enlargement in the lungs of mice lacking tissue inhibitor of metalloproteinases-3 (TIMP-3). *J. Clin. Invest.* **108**, 817–829 (2001).
19. Kasahara, Y. *et al.* Inhibition of VEGF receptors causes lung cell apoptosis and emphysema. *J. Clin. Invest.* **106**, 1311–1319 (2000).
20. Niewoehner, D. E., Kleinerman, J. & Rice, D. B. Pathologic changes in the peripheral airways of young cigarette smokers. *N. Engl. J. Med.* **291**, 755–758 (1974).
21. Silverman, E. K. *et al.* Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. *Am. J. Respir. Crit. Care Med.* **157**, 1770–1778 (1998).
22. Wikenheiser, K. A., Clark, J. C., Linnoila, R. I., Stahlman, M. T. & Whitsett, J. A. Simian virus 40 large T antigen directed by transcriptional elements of the human surfactant protein C gene produces pulmonary adenocarcinomas in transgenic mice. *Cancer Res.* **52**, 5342–5352 (1992).
23. Cone, R. L., Weinacker, A., Chen, A. & Sheppard, D. Effects of beta subunit cytoplasmic domain deletions on the recruitment of the integrin $\alpha v \beta 6$ to focal contacts. *Cell Adhes. Commun.* **2**, 101–113 (1994).
24. Clark, J. C. *et al.* FGF-10 disrupts lung morphogenesis and causes pulmonary adenomas *in vivo*. *Am. J. Physiol. Lung. Cell. Mol. Physiol.* **280**, L705–L715 (2001).
25. Liu, X. *et al.* Conditional epidermal expression of TGF β 1 blocks neonatal lethality but causes a reversible hyperplasia and alopecia. *Proc. Natl Acad. Sci. USA* **98**, 9139–9144 (2001).
26. Dolganov, G. M. *et al.* A novel method of gene transcript profiling in airway biopsy homogenates reveals increased expression of a Na⁺-K⁺-Cl⁻ cotransporter (NKCC1) in asthmatic subjects. *Genome Res.* **11**, 1473–1483 (2001).
27. Dunnill, M. S. Quantitative methods in the study of pulmonary pathology. *Thorax* **17**, 320–328 (1962).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements This work was supported by grants from the NHLBI to D.G.M. and D.S., including a Program for Genomic Applications Grant (Baygenomics).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.S. (e-mail: deans@itsa.ucsf.edu).

Free fatty acids regulate insulin secretion from pancreatic β cells through GPR40

Yasuaki Itoh^{*†}, Yuji Kawamata^{*†}, Masataka Harada^{*}, Makoto Kobayashi^{*}, Ryo Fujii^{*}, Shoji Fukusumi^{*}, Kazuhiro Ogi^{*}, Masaki Hosoya^{*}, Yasuhiro Tanaka^{*}, Hiroshi Uejima^{*}, Hideyuki Tanaka^{*}, Minoru Maruyama^{*}, Rie Satoh^{*}, Shoichi Okubo^{*}, Hideki Kizawa^{*}, Hidetoshi Komatsu^{*}, Fumika Matsumura^{*}, Yuko Noguchi^{*}, Tokuyuki Shinohara^{*}, Shuji Hinuma^{*}, Yukio Fujisawa^{*} & Masahiko Fujino^{*}

^{*} Discovery Research Laboratories I, Pharmaceutical Research Division, Takeda Chemical Industries, Ltd, Wadai 10, Tsukuba, Ibaraki 300-4293, Japan

[†] These authors contributed equally to this work

Diabetes, a disease in which carbohydrate and lipid metabolism are regulated improperly by insulin, is a serious worldwide health issue^{1,2}. Insulin is secreted from pancreatic β cells in response to elevated plasma glucose, with various factors modifying its secretion³. Free fatty acids (FFAs) provide an important energy source as nutrients, and they also act as signalling molecules in various cellular processes, including insulin secretion^{4,5}. Although FFAs are thought to promote insulin secretion in an acute phase, this mechanism is not clearly understood⁶. Here we show that a G-protein-coupled receptor, GPR40,

which is abundantly expressed in the pancreas, functions as a receptor for long-chain FFAs. Furthermore, we show that long-chain FFAs amplify glucose-stimulated insulin secretion from pancreatic β cells by activating GPR40. Our results indicate that GPR40 agonists and/or antagonists show potential for the development of new anti-diabetic drugs.

GPR40 is an orphan (that is, its ligands are unidentified) G-protein-coupled receptor (GPCR) isolated originally from a human genomic DNA fragment⁷. We isolated complementary DNAs from human, monkey, mouse, rat and hamster (DDBJ/EMBL/GenBank accession numbers AF024687, AB095743, AB095744, AB095745 and AB095746, respectively) and found that their amino acid sequences (300 amino acids) were highly conserved. Analyses for the distribution of GPR40 messenger RNA in rat tissue by quantitative polymerase chain reaction with reverse transcription (RT-PCR)⁸ demonstrated its highest expression level in the pancreas (Fig. 1a). In addition, we found that expression of GPR40 mRNA

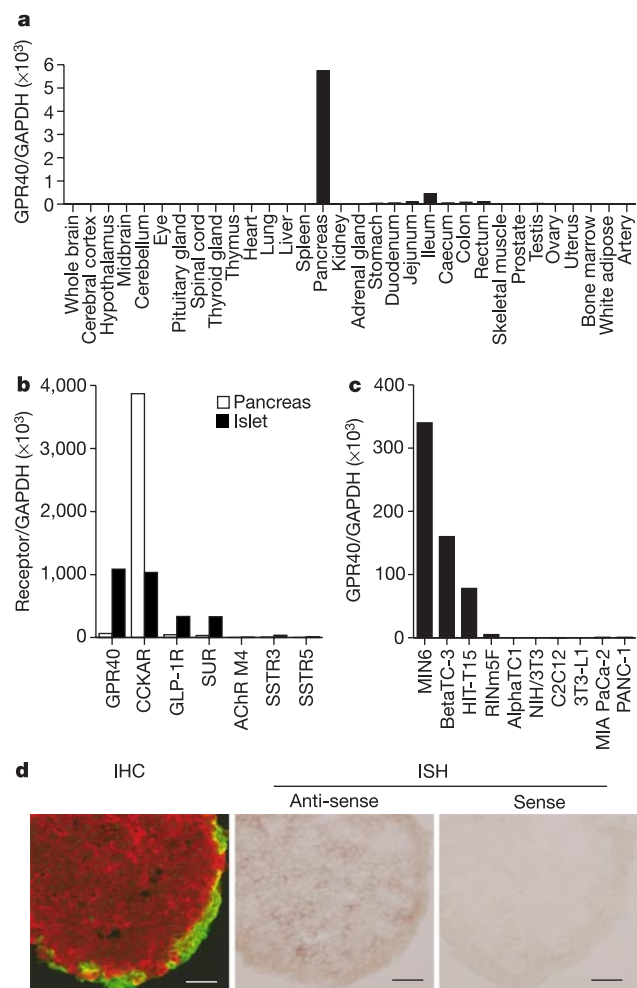


Figure 1 GPR40 mRNA is expressed abundantly in pancreatic β cells. Data represent the ratios of GPR40 and other mRNAs to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA. **a**, Distribution of GPR40 mRNA in rat tissues. **b**, Specific expression of GPR40 mRNA in rat pancreatic islets. CCKAR, type A cholecystokinin receptor; SSTR, somatostatin receptor; SUR, sulphonylurea receptor. **c**, Expression of GPR40 mRNA in pancreatic β -cell lines. MIN6, mouse pancreatic β cells; betaTC-3, mouse pancreatic β cells; HIT-T15, Syrian golden hamster pancreatic β cells; RINm5F, rat pancreatic β cells; alphaTC1, mouse pancreatic α cells; NIH/3T3, mouse embryonic cells; C2C12, mouse myoblasts; 3T3-L1, mouse embryonic fibroblasts; MIA PaCa-2, human pancreatic carcinoma cells; PANC-1, human pancreatic carcinoma cells. **d**, Localization of GPR40 mRNA in rat islet cells. IHC, immunohistochemistry; ISH, *in situ* hybridization. Scale bar, 50 μ m.

was approximately 17 times greater in the pancreatic islets than in the pancreas as a whole (Fig. 1b). Its expression was comparable to that of type A cholecystokinin receptor⁹, glucagon-like peptide-1 receptor (GLP-1R)¹⁰ and sulphonylurea receptor¹¹, and greater than that of muscarinic acetylcholine receptor (AChR M4)¹² and somatostatin receptors 3 and 5 (ref. 13), all of which are reported to be expressed abundantly in the islets. Furthermore, GPR40 mRNA was found to be expressed significantly in pancreatic β -cell lines, with its highest level detected in MIN6 cells (Fig. 1c)¹⁴. We found no evidence of GPR40 mRNA expression in any other cell lines including a pancreatic α -cell line. Furthermore, we performed double staining with *in situ* hybridization for GPR40 mRNA and immunohistochemistry for insulin or glucagon in rat islets. We mainly detected positive signals in areas stained by anti-insulin antibody (β cells; Fig. 1d). We did not, however, detect clear signals in areas stained by anti-glucagon antibody (α cells). Thus, GPR40 seems to be expressed mainly in β cells rather than α cells, although future studies are necessary to determine the precise expression levels of GPR40 in these cells.

Previously, we established a strategy to identify ligands by detecting signal transductions in cells expressing orphan GPCRs^{15,16}. To apply this technique to the identification of GPR40 ligands, we transiently expressed human GPR40 cDNA in Chinese hamster ovary (CHO) cells, added over 1,000 chemical compounds, and then examined changes in intracellular Ca^{2+} concentration ($[\text{Ca}^{2+}]_i$) with a fluorometric imaging plate reader (FLIPR) system⁸. As a result, long-chain FFAs were found to evoke a specific rise of $[\text{Ca}^{2+}]_i$ in CHO cells expressing GPR40 (data not shown). Next, we prepared CHO cells stably expressing human, mouse or rat GPR40 (CHO-hGPR40, CHO-mGPR40 and CHO-rGPR40, respectively), and determined the Ca^{2+} influx-inducing activities of various FFAs on these cells (Fig. 2a). Apparent stimulatory activities were detected in saturated FFAs of C12 to C16 length and in both C18- and C20-length unsaturated FFAs (Table 1). However, methyl linoleate did not show stimulatory activity, suggesting that the carboxyl group is indispensable for this function. Some eicosanoids

also showed stimulatory activities on CHO-hGPR40 cells in levels comparable to long-chain FFAs.

Although we found that long-chain FFAs slightly decreased 3',5'-cyclic AMP (cAMP) production in forskolin-stimulated CHO-hGPR40 cells (data not shown), the FFA-induced increase of $[\text{Ca}^{2+}]_i$ in CHO-hGPR40 cells was not affected by treatment with pertussis toxin (data not shown), suggesting that human GPR40 couples mainly with a G-protein α -subunit of the G_q family ($G_{q\alpha}$) and partially with $G_{i\alpha}$. However, we could not detect changes in cAMP production in CHO-mGPR40 cells (data not shown), suggesting that there is a species difference in the coupling of GPR40 with G proteins, and that mouse GPR40 couples with $G_{q\alpha}$ only in CHO cells. We subsequently examined extracellular signal-regulated kinase mitogen-activated protein (MAP) kinase in CHO-hGPR40 cells, as the stimulation of GPCRs frequently leads to MAP kinase activation. The long-chain FFAs, oleic acid, linoleic acid and docosahexaenoic acid (DHA), increased phosphorylated MAP kinase (p44/42) in CHO-hGPR40 cells, but not in mock CHO cells (transfected with an expression vector not containing GPR40 cDNA) (Fig. 2b). Methyl linoleate did not affect MAP kinase activation in either the mock CHO cells or CHO-hGPR40 cells. These results demonstrate that GPR40 functions as a specific receptor for long-chain FFAs.

Although it has been reported that FFAs activate GPR40 and that GPR40 is highly expressed in pancreatic β cells¹⁷, its function remains unknown. We found that oleic acid, linoleic acid, γ -linolenic acid, arachidonic acid and DHA induced a $[\text{Ca}^{2+}]_i$ increase in MIN6 cells, whereas methyl linoleate and butyric acid did not (Fig. 3a). Changes in cAMP production were not detected in MIN6 cells treated with FFAs (data not shown), suggesting that GPR40 couples with $G_{q\alpha}$ even in MIN6 cells. Furthermore, we detected FFA-induced MAP kinase activation in MIN6 cells similar to CHO-hGPR40 cells (Fig. 3b). We subsequently examined the effects of FFAs on insulin secretion. Oleic acid, linoleic acid, α -linolenic acid, γ -linolenic acid, arachidonic acid and DHA stimulated insulin secretion in MIN6 cells, but methyl linoleate and butyric acid did not (Fig. 3c). Thus, FFAs with potent stimulatory activity on CHO-mGPR40 cells appeared to show insulin-secretion-inducing activity on MIN6 cells. The stimulatory activities of oleic and linoleic acids on insulin secretion were detected more strongly in high-glucose than in low-glucose concentrations (Fig. 3d), indicating that FFAs amplify glucose-stimulated insulin secretion (GSIS) from MIN6 cells.

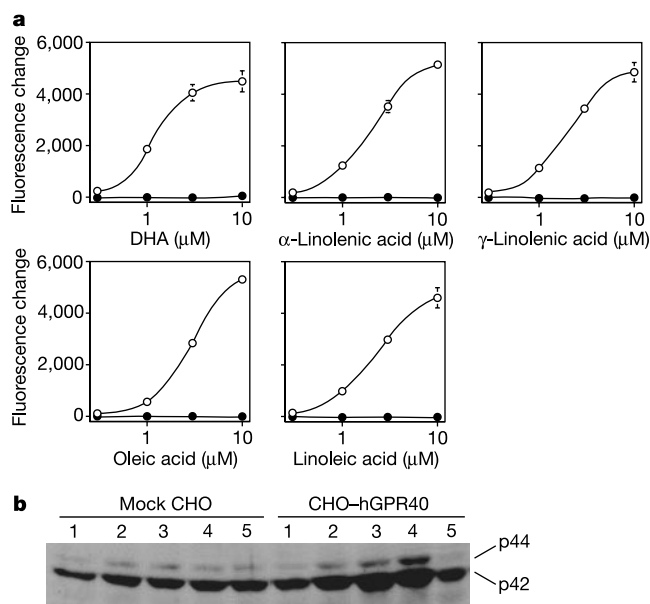


Figure 2 Induction of $[\text{Ca}^{2+}]_i$ rise and MAP kinase activation by FFAs in CHO-hGPR40 cells. **a**, Representative dose-response curves of FFA-induced $[\text{Ca}^{2+}]_i$ rise in CHO-hGPR40 cells. Open circles, CHO-hGPR40 cells; filled circles, control cells (CHO cells expressing human histamine H1 receptor). Data represent the mean values $\pm$ s.e.m. of the maximal fluorescent changes induced by FFAs in three separate experiments with a FLIPR assay. **b**, MAP kinase activation in CHO-hGPR40 cells induced by FFAs. Lanes: 1, no FFA; 2, linoleic acid; 3, oleic acid; 4, DHA; 5, methyl linoleate. Treatments were for 10 min.

Table 1 Fatty acid potency to induce $[\text{Ca}^{2+}]_i$ rise in CHO cells expressing GPR40

FFA	(EC ₅₀ , μM)		
	Human	Mouse	Rat
Acetic acid (C2)	Inactive	Inactive	Inactive
Butyric acid (C4)	Inactive	Inactive	Inactive
Caproic acid (C6)	Inactive	Inactive	>300
Caprylic acid (C8)	>300	Inactive	>300
Capric acid (C10)	43 $\pm$ 2.2	>100	>100
Lauric acid (C12)	5.7 $\pm$ 1.4	5.6 $\pm$ 1.6	13 $\pm$ 3.3
Myristic acid (C14)	7.7 $\pm$ 1.4	6.0 $\pm$ 0.8	7.3 $\pm$ 0.5
Palmitic acid (C16)	6.8 $\pm$ 0.5	4.6 $\pm$ 1.2	6.6 $\pm$ 0.4
Stearic acid (C18)	>300	>300	>300
Oleic acid (C18:1)	2.0 $\pm$ 0.3	2.7 $\pm$ 0.5	3.4 $\pm$ 0.4
Elaidic acid (C18:1)	4.7 $\pm$ 0.4	6.5 $\pm$ 1.5	11 $\pm$ 1.5
Linoleic acid (C18:2)	1.8 $\pm$ 0.1	2.9 $\pm$ 0.3	4.1 $\pm$ 0.5
Methyl linoleate	Inactive	Inactive	Inactive
α -Linolenic acid (C18:3)	2.0 $\pm$ 0.3	3.6 $\pm$ 0.3	4.0 $\pm$ 0.7
γ -Linolenic acid (C18:3)	4.6 $\pm$ 1.6	5.2 $\pm$ 0.6	5.4 $\pm$ 1.1
Arachidonic acid (C20:4)	2.4 $\pm$ 0.6	5.4 $\pm$ 0.8	8.0 $\pm$ 0.6
Eicosapentaenoic acid (C20:5)	2.3 $\pm$ 0.4	4.9 $\pm$ 0.8	9.8 $\pm$ 0.6
Docosahexaenoic acid (C22:6)	1.1 $\pm$ 0.3	16 $\pm$ 4.7	13 $\pm$ 1.7
5,6-Epoxyeicosatrienoic acid	7.7 $\pm$ 0.2	ND	ND
8,9-Epoxyeicosatrienoic acid	6.1 $\pm$ 0.2	ND	ND
11,12-Epoxyeicosatrienoic acid	1.4 $\pm$ 0.3	ND	ND
14,15-Dihydroxyeicosatrienoic acid	1.1 $\pm$ 0.4	ND	ND

EC₅₀ indicates the concentration of a sample that produces 50% of the maximal response, and was calculated from dose-response curves. Data represent the mean values $\pm$ s.e.m. in 3–6 assays. Inactive, no response at 300 μM ; ND, not done.

We next performed experiments to inhibit the expression of GPR40 in MIN6 cells by using small interfering RNA (siRNA). The increase of insulin secretion from MIN6 cells after stimulation with linoleic acid and γ -linolenic acid was apparently eliminated by treatment with a siRNA specific for mouse GPR40, whereas the increase of insulin secretion after stimulation with GLP-1 was unaffected (Fig. 3e). We confirmed that the treatment with siRNA reduced GPR40 mRNA expression by more than half, but had no effect on GLP-1R mRNA expression (Fig. 3e). These results indicate that at least a part of the FFA-amplified GSIS from MIN6 cells is mediated through GPR40. Furthermore, we demonstrated that linoleic acid and γ -linolenic acid induced insulin secretion from normal rat islets as well as MIN6 cells (Fig. 3f).

Levels of FFAs in plasma are usually 0.2–1.7 mM, but 99% or more of these FFAs are bound tightly with serum albumin so that the concentration of unbound FFAs is in the range of 0.01–10 μ M¹⁸. We found that bovine serum albumin (BSA) at more than 0.1% inhibited the Ca^{2+} -influx-inducing activities of FFAs on

CHO–hGPR40 cells (Fig. 3g). Furthermore, the presence of 0.1% BSA also obstructed FFA-induced insulin secretion from MIN6 cells (Fig. 3h). These results suggest that FFAs unbound to albumin act to stimulate GPR40. FFAs have been reported to modulate insulin secretion^{6,19}: in short-term exposure, FFAs enhance GSIS but under long-term exposure they act to attenuate it; an effect called lipotoxicity²⁰. Although several studies have proposed possible mechanisms for FFA-amplified GSIS—for example, that FFAs might simply provide energy to β cells or, alternatively, that the metabolites of FFAs, fatty acyl-coenzyme A molecules, act to stimulate insulin secretion^{5,18,21}—our results indicate that GPR40 takes an active part in FFA-amplified GSIS.

Increases in $[\text{Ca}^{2+}]_i$ as well as cAMP concentration have important roles in triggering insulin secretion²². We found that depriving the assay medium of Ca^{2+} eliminated FFA-amplified GSIS from MIN6 cells (Fig. 3i). In addition, an L-type calcium channel blocker, nifedipine, inhibited GSIS from MIN6 cells treated with FFAs (Fig. 3j), suggesting that Ca^{2+} influx into cells from the extracellular

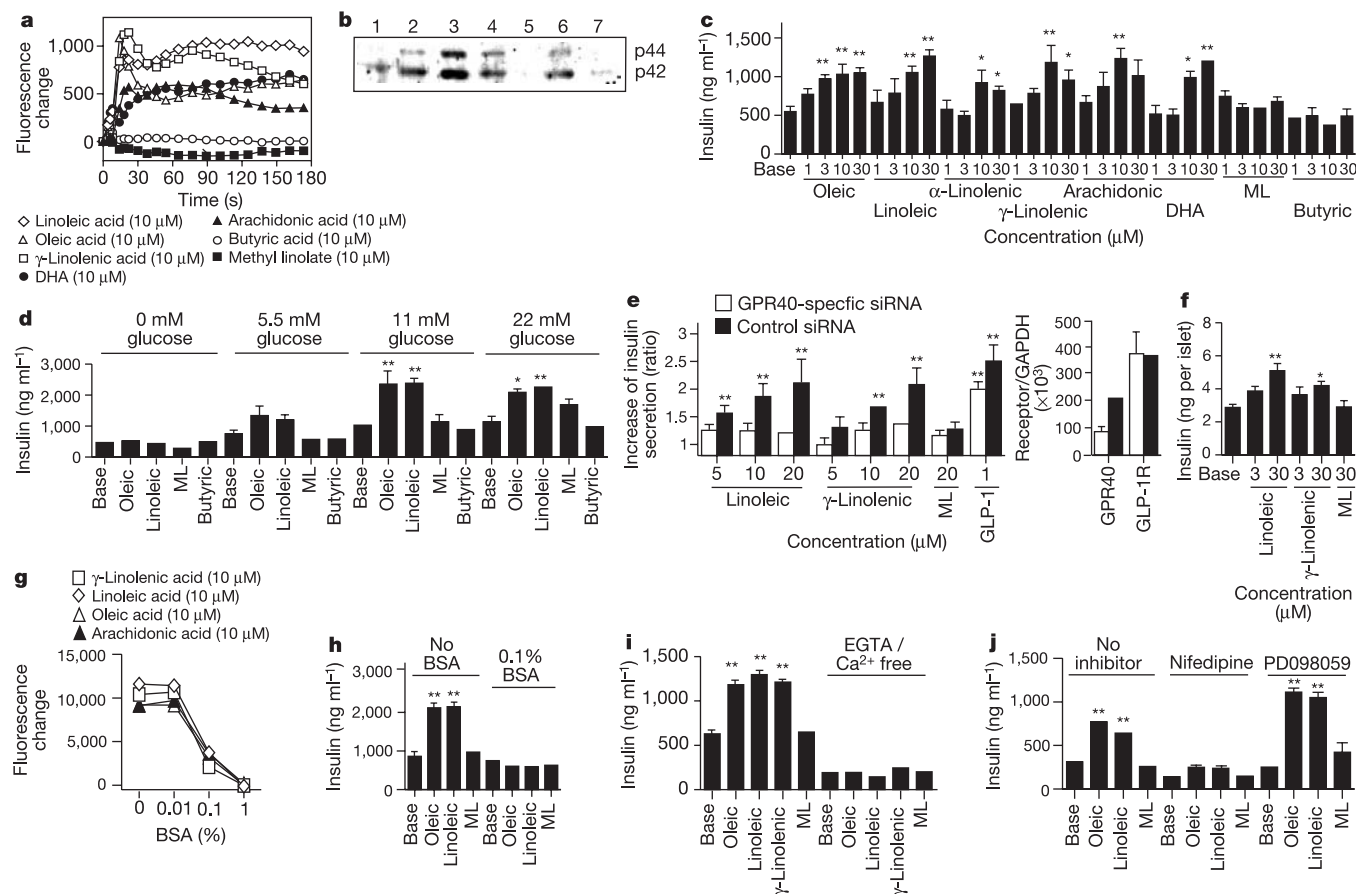


Figure 3 Analyses for FFA-induced insulin secretion mechanisms. Asterisk, $P < 0.05$; double asterisk, $P < 0.01$ (Student's t -test). **a**, FFA-induced $[\text{Ca}^{2+}]_i$ rise in MIN6 cells. Each trace represents the mean value in duplicate assays with a FLIPR. **b**, FFA-activated MAP kinase in MIN6 cells. Lanes: 1, no FFA; 2, oleic acid; 3, linoleic acid; 4, γ -linolenic acid; 5, methyl linoleate (all treatments 2.5 min); 6, γ -linolenic acid; 7, γ -linolenic acid and 50 μ M PD98059 (all treatments 5 min). **c**, FFA-amplified insulin secretion from MIN6 cells. Data represent the mean $\pm$ s.e.m. in quadruplicate assays in the presence of 22 mM glucose. ML, methyl linoleate. **d**, FFA-induced GSIS from MIN6. Data represent the mean $\pm$ s.e.m. in quadruplicate assays (10 μ M FFAs). **e**, Effect of siRNA on FFA-induced increase of insulin secretion and on expression of GPR40 mRNA in MIN6 cells. Data for the increase of insulin secretion (left) represent the mean $\pm$ s.e.m. as ratios of samples compared to the basal level in two independent experiments ($n = 4$ in each experiment). Basal levels were 59.6 and 53.0 ng ml⁻¹ after treatment with GPR40-specific and control siRNAs, respectively. Data on mRNA expression (right)

represent the ratios of GPR40 and GLP-1R mRNAs to GAPDH mRNA in duplicate assays with RT-PCR. **f**, FFA-amplified insulin secretion from primary cultured rat islets. Data represent the mean $\pm$ s.e.m. in two separate experiments ($n = 3$ and 4, respectively). **g**, Effects of BSA on the FFA-induced $[\text{Ca}^{2+}]_i$ rise in CHO–hGPR40 cells. Each trace represents the mean value in duplicate assays. **h**, Effects of BSA on FFA-amplified GSIS from MIN6 cells. Pre-incubation and treatment buffers: KRBH in the absence (left) or presence (right) of 0.1% fatty-acid-free BSA; treatment: FFAs (10 μ M) with 22 mM glucose. Data represent the mean $\pm$ s.e.m. in quadruplicate assays. **i**, Role of Ca^{2+} in FFA-amplified GSIS from MIN6 cells. Pre-incubation and treatment buffers: KRBH (left) or Ca^{2+} -free KRBH containing 0.1 mM EGTA (right); treatment: FFAs (10 μ M) with 22 mM glucose. Data represent the mean $\pm$ s.e.m. in quadruplicate assays. **j**, Effects of inhibitors on FFA-amplified insulin secretion from MIN6 cells. Pre-incubation and treatment buffers: KRBH (left), KRBH containing 10 μ M nifedipine (middle) or 50 μ M PD098059 (right). FFA treatment and data representation are the same as **i**.

environment is critical for FFA-amplified GSIS from these cells. However, in FLIPR assays, nifedipine had no effect on the rapid rise of $[Ca^{2+}]_i$ in MIN6 cells induced by FFAs (data not shown). This result indicates that the $[Ca^{2+}]_i$ increase indicated in Fig. 3a was due to Ca^{2+} influx through nifedipine-insensitive calcium channels, including Ca^{2+} release from intracellular Ca^{2+} stores such as the endoplasmic reticulum. We believe that the rapid $[Ca^{2+}]_i$ rise induced by FFAs would act to amplify the subsequent glucose-dependent $[Ca^{2+}]_i$ rise that is sensitive to nifedipine and essential for GSIS, because this sort of contribution to the enhancement of GSIS by GPCR signalling in pancreatic β cells has been reported for ACHR²³. On the other hand, we found that a MAP kinase inhibitor, PD098059, had no effect on FFA-amplified GSIS from MIN6 cells (Fig. 3j) under conditions where it almost completely inhibited the activation of MAP kinase (Fig. 3b). This suggests that the activation of MAP kinase is not necessary for insulin secretion from MIN6 cells. As the MAP kinase pathway is closely linked to growth and differentiation signals²⁴, the stimulatory effects of FFAs through GPR40 might evoke other cellular events aside from insulin secretion. Although our data indicate that FFAs act as ligands for GPR40, binding experiments using proper synthetic compounds with high affinity are required to analyse the direct interaction between FFAs and GPR40. We believe that our results will provide new insight into insulin secretion mechanisms and the treatment of diabetes. □

Methods

Histochemistry

We isolated pancreatic islets from adult male Wistar rats²⁵. Fresh frozen sections (16 μ m) were attached to silanized slides and fixed with 4% paraformaldehyde. *In situ* hybridization was carried out using digoxigenin-labelled riboprobes. Rabbit anti-insulin and monoclonal anti-glucagon antibodies were used for the double staining of *in situ* hybridization and immunohistochemistry, and insulin- and glucagon-positive signals were visualized with Alexa488- or Alexa594-labelled second antibodies, respectively.

MAP kinase

MAP kinase activation in CHO and MIN6 cells after treatment with 10 μ M of each FFA was determined as described previously²⁶. MIN6 cells (donated by J. Miyazaki) were cultured in DMEM medium containing 25 mM glucose, 15% fetal bovine serum, 5.5 μ M 2-mercaptoethanol, 20 mM HEPES, pH 7.3, and antibiotics for 2 days before the assay.

Insulin secretion

MIN6 cells were cultured in the medium for 2 days and then washed with a modified Krebs–Ringer bicarbonate buffer (KRBH). After pre-incubation in KRBH either with or without reagents, the cells were treated for 90 min with FFAs dissolved in KRBH supplemented with glucose. Normal islets were isolated from male Wistar rats (8 weeks, Charles River Japan) by digesting the pancreas with collagenase and then washing with KRBH containing 2.8 mM glucose. Ten to fourteen islets were pooled together and pre-incubated with the same buffer, then treated for 90 min with FFAs dissolved in KRBH supplemented with 16.5 mM glucose. The amount of insulin in the culture supernatant was determined with a RIA or EIA kit (Amersham Pharmacia).

siRNA

Five siRNAs were designed from the mouse GPR40 sequence. These siRNAs, containing 21 nucleotides, were obtained from Dharmacon Research. In our experiments to suppress GPR40 expression in MIN6 cells, we used the most effective siRNAs from among them (target sequences of 5'-CGCCAGUUGACAUUCUdTdT-3' and 5'-AAGAAUGU CACAACUGGCGdTdT-3'). As a control, we used scramble II duplex siRNA (Dharmacon Research). Transfection of siRNA was accomplished with an HVJ envelope Vector Kit GenomONE (Ishihara Sangyo Kaisha). Two days after transfection, the cells were treated with FFAs and insulin secretion and GPR40 mRNA expression were determined.

Received 2 January; accepted 11 February 2003; doi:10.1038/nature01478.

Published online 23 February 2003.

- Gadsby, R. Epidemiology of diabetes. *Adv. Drug Deliv. Rev.* **54**, 1165–1172 (2002).
- Jovanovic, L. & Gondon, B. Type 2 diabetes: The epidemic of the new millennium. *Ann. Clin. Lab. Sci.* **29**, 33–42 (1999).
- Henquin, J. C. Triggering and amplifying pathways of regulation of insulin secretion by glucose. *Diabetes* **49**, 1751–1760 (2000).
- Nunez, E. A. Biological complexity is under the 'strange attraction' of non-esterified fatty acids. *Prostaglandins Leukot. Essent. Fatty Acids* **57**, 107–110 (1997).
- Haber, E. P. et al. Pleiotropic effects of fatty acids on pancreatic β -cells. *J. Cell Physiol.* **194**, 1–12 (2003).
- Dobbins, R. L. et al. A fatty acid-dependent step is critically important for both glucose- and non-glucose-stimulated insulin secretion. *J. Clin. Invest.* **101**, 2370–2376 (1998).
- Sawzargo, M. et al. A cluster of four novel human G protein-coupled receptor genes occurring in

- close proximity to CD22 gene on chromosome 19q13.1. *Biochem. Biophys. Res. Commun.* **239**, 543–547 (1997).
- Fujii, R. et al. Identification of neuromedin U as the cognate ligand of the orphan G protein-coupled receptor FM-3. *J. Biol. Chem.* **275**, 21068–21074 (2000).
- Wank, S. A. et al. Purification, molecular cloning, and functional expression of the cholecystokinin receptor from rat pancreas. *Proc. Natl Acad. Sci. USA* **89**, 3125–3129 (1992).
- Lankat-Buttgereit, B., Goke, R., Fehmann, H. C., Richter, G. & Goke, B. Molecular cloning of a cDNA encoding for the GLP-1 receptor expressed in rat lung. *Exp. Clin. Endocrinol.* **102**, 341–347 (1994).
- Aguilar-Bryan, L. et al. Cloning of the beta cell high-affinity sulfonylurea receptor: a regulator of insulin secretion. *Science* **268**, 423–426 (1995).
- Braun, T., Schofield, P. R., Shivers, B. D., Pritchett, D. B. & Seeburg, P. H. A novel subtype of muscarinic receptor identified by homology screening. *Biochem. Biophys. Res. Commun.* **149**, 125–132 (1987).
- Bruno, J. F., Xu, Y., Song, J. & Berelowitz, M. Tissue distribution of somatostatin receptor subtype messenger ribonucleic acid in the rat. *Endocrinology* **133**, 2561–2567 (1993).
- Miyazaki, J. et al. Establishment of a pancreatic beta cell line that retains glucose-inducible insulin secretion: special reference to expression of glucose transporter isoforms. *Endocrinology* **127**, 126–132 (1990).
- Hinuma, S. et al. A prolactin-releasing peptide in the brain. *Nature* **393**, 272–276 (1998).
- Hinuma, S., Onda, H. & Fujino, M. The quest for novel bioactive peptides utilizing orphan seven-transmembrane-domain receptors. *J. Mol. Med.* **77**, 495–504 (1999).
- Briscoe, C. P. et al. The orphan G protein-coupled receptor GPR40 is activated by medium and long-chain fatty acids. *J. Biol. Chem.* published online, 19 December 2002 (doi:211495200).
- Spector, A. A. & Hoak, J. C. Fatty acids, platelets, and microcirculatory obstruction. *Science* **190**, 490–492 (1975).
- Stein, D. T. et al. Essentiality of circulating fatty acids for glucose-stimulated insulin secretion in the fasted rat. *J. Clin. Invest.* **97**, 2728–2735 (1996).
- Bergman, R. N. & Ader, M. Free fatty acids and pathogenesis of type 2 diabetes mellitus. *Trends Endocrinol. Metab.* **11**, 351–356 (2000).
- Newgard, C. B. & McGarry, J. D. Metabolic coupling factors in pancreatic β -cell signal transduction. *Annu. Rev. Biochem.* **64**, 689–719 (1995).
- Ashcroft, F. M. et al. Stimulus-secretion coupling in pancreatic beta cells. *J. Cell Biochem.* **55**, 54–65 (1994).
- Wolheim, C. B. & Biden, T. J. Signal transduction in insulin secretion: comparison between fuel stimuli and receptor agonists. *Ann. NY Acad. Sci.* **488**, 317–333 (1986).
- Gutkind, J. S. Cell growth control by G protein-coupled receptors: from signal transduction to signal integration. *Oncogene* **17**, 1331–1342 (1998).
- Lacy, P. E. & Kostianovsky, M. Method for the isolation of intact islets of Langerhans from the rat pancreas. *Diabetes* **16**, 35–39 (1967).
- Kawamata, Y. et al. A G protein-coupled receptor responsive to bile acids. *J. Biol. Chem.* published online, 10 January 2003 (doi:209706200).

Acknowledgements We thank Y. Sumino, T. Soda, K. Kato, H. Odaka, F. Itoh, M. Kajino, K. Fukatsu and N. Suzuki for their discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.H. (e-mail: Hinuma_Shuji@takeda.co.jp).

Functional and spatial segregation of secretory vesicle pools according to vesicle age

Rory R. Duncan*, Jennifer Greaves*, Ulrich K. Wiegand*, Ioulia Matskevich*, Georg Bodammer*†, David K. Apps*, Michael J. Shipston* & Robert H. Chow‡

* Membrane Biology Group, University of Edinburgh, George Square, Edinburgh EH8 9XD, UK

‡ Department of Physiology and Biophysics, Keck USC School of Medicine, Los Angeles, California 90089-9142, USA

Synaptic terminals and neuroendocrine cells are packed with secretory vesicles, only a few of which are docked at the plasma membrane and readily releasable. The remainder are thought to constitute a large cytoplasmic reserve pool awaiting recruitment into the readily releasable pool (RRP) for exocytosis¹. How vesicles are prioritized in recruitment is still unknown: the choice

† Present address: MicroEmissive Displays Ltd, Scottish Microelectronics Centre, West Mains Road, Edinburgh EH9 3JF, UK.

could be random, or else the oldest or the newest ones might be favoured. Here we show, using a fluorescent cargo protein² that changes colour with time³, that vesicles in bovine adrenal chromaffin cells segregate into distinct populations, based on age. Newly assembled vesicles are immobile (morphologically docked) at the plasma membrane shortly after biogenesis, whereas older vesicles are mobile and located deeper in the cell. Different secretagogues selectively release vesicles from the RRP or, surprisingly, selectively from the deeper cytoplasmic pool. Thus, far from being equal, vesicles are segregated functionally and spatially according to age.

Bovine adrenal chromaffin cells (BCCs) package atrial natriuretic factor (ANF) with catecholamines in large dense-core vesicles

(LDCVs)⁴. To investigate the fate of newly assembled LDCVs, we targeted fluorescent ANF fusion proteins to the vesicle lumen. After 48 h of expression of a fusion of ANF with enhanced green fluorescent protein (ANF-EGFP) there were 157 ± 17 (mean $\pm$ s.e.m.) punctate fluorescent objects per cell ($n = 12$ cells from eight cell preparations), concentrated in the lower half of the cells (that is, towards the substrate; see Supplementary Fig. 1) with a mean diameter of 380 nm ($n = 619$ objects; the s.d. was below the limits of optical resolution), as has previously been noted for LDCVs in BCCs^{5,6}. ANF-EGFP co-localized ($96.5 \pm 0.2\%$) with LysoTracker Red DND-99 in LDCV-sized structures, indicating that virtually all ANF-EGFP was localized to acidic structures, as expected for targeting to LDCVs (Supplementary Fig. 1a).

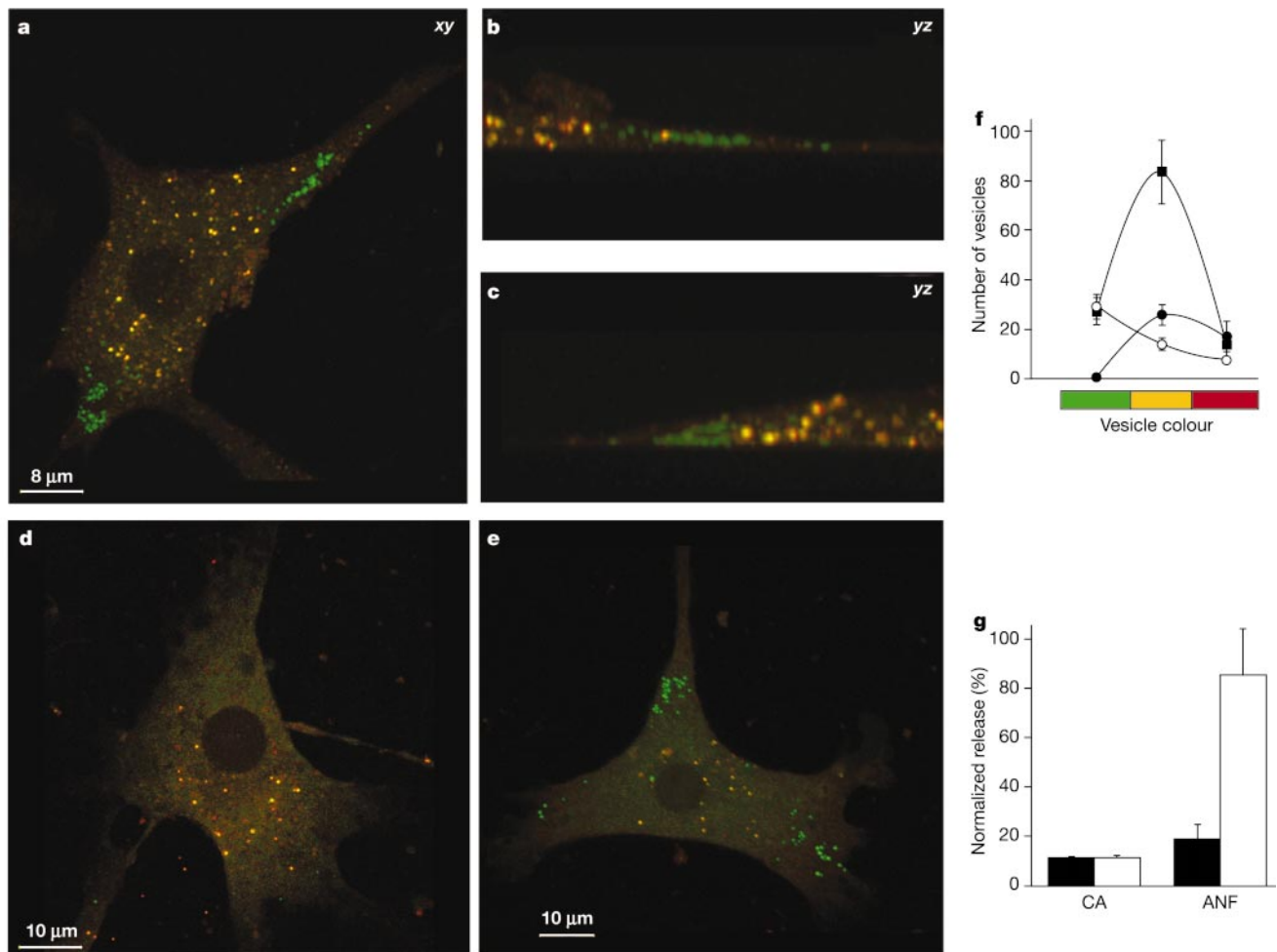


Figure 1 ANF-dsRed-E5-transduced chromaffin cells contain populations of LDCVs with distinct colours and distributions. **a**, A representative, projected image of 48, 0.2 μ m-thick confocal sections, showing a control cell with peripheral green LDCVs and evenly distributed yellow and red vesicles. **b, c**, All data were deconvolved, increasing the apparent z-plane resolution of the images. This allowed a three-dimensional reconstruction of two different regions of the cell in **a** to be viewed from the yz direction, thus revealing that the green LDCVs are almost exclusively membrane-proximal. **d**, Projected image of 48, 0.2 μ m-thick confocal sections, showing that the numbers of green and yellow fluorescently labelled LDCVs are significantly decreased after a maximal nicotinic stimulation. **e**, Projected image of 24, 0.2 μ m-thick confocal sections, showing that the number of yellow fluorescently labelled LDCVs is significantly decreased after maximal stimulation with barium, but that morphologically docked green LDCVs remain. **f**, These data are emphasized by plotting the mean number of vesicles per cell before (filled squares) and after (filled circles) maximal nicotinic stimulation or maximal barium stimulation (open circles) in each population, illustrating the near-total loss of green vesicles after nicotinic stimulation but not under barium stimulation: there was 65–85%

loss of yellow vesicles and no significant decrease in the number of the oldest (that is, red) vesicles. **g**, Population assays of exocytosis in control and transduced cells. A histogram illustrating the fractional catecholamine (CA) and ANF release from control (filled bars) and transduced cells (open bars) after maximal stimulation. Expression of ANF-EGFP or ANF-dsRed-E5 in cultured BCCs did not significantly affect the cellular catecholamine content (73 ± 8 and 52 ± 7 fmol per cell, $n = 51$ and 22 , respectively, $P > 0.27$; see also ref. 6) at 48–72 h after viral transduction. Nor was maximal catecholamine secretion affected: control cells secreted $15.5 \pm 2.7\%$ of total cellular catecholamine ($n = 31$), whereas transduced cells secreted $14.4 \pm 2.4\%$ ($n = 31$, $P > 0.99$). The ANF content of control cells was found to be 99 ± 12 fmol per 10^6 cells, not significantly different from the published value of 150 fmol per 10^6 cells (ref. 4). Of this, $18.8 \pm 6.1\%$ ($n = 15$) was released, a fraction not significantly different from the fraction of catecholamine released ($P > 0.06$). Transduced BCCs overexpressed ANF about threefold (335 ± 45 fmol per 10^6 cells); of this, $85.4 \pm 18.5\%$ was released ($n = 12$). Stimulation with nicotine in the absence of external Ca^{2+} did not elicit a significant release of catecholamine (not shown).

Analysis of the same cells before and after maximal stimulation (20 μ M nicotine, 15 min) showed that the number of ANF-EGFP-labelled vesicles fell from 157 ± 17 to 63 ± 3 ($n = 12$ cells) per cell, representing a loss of $60 \pm 7\%$ of labelled LDCVs (Supplementary Fig. 1b, c; $P = 0.001$). This fraction is significantly greater than expected if older vesicles were given priority (few fluorescent vesicles should have been released in this case, because all fluorescent vesicles were less than 48 h old, much less than the estimated vesicle lifetime of 18 days⁷) or if recruitment were random (in which case the fraction of vesicles released should approximate the fraction of stored catecholamine released, that is, 10–15% (refs 8–10)). ANF-EGFP vesicles remaining after maximal nicotine stimulation were mostly centrally distributed; that is, peripheral vesicles had disappeared (Supplementary Fig. 1c). The disproportionately large fractional release of peripheral vesicles less than 48 h old suggests that the most recently assembled LDCVs are quickly mobilized to the plasma membrane and preferentially recruited for secretion by nicotine in BCCs. This analysis does not exclude the possibility that vesicles at the plasma membrane might be withdrawn into the cell rather than undergo exocytosis; however, the marked decrease in the number of vesicles and the constant size of spared vesicles make this unlikely.

To further refine our understanding of how vesicles of different ages are selected for secretion, we labelled them with a fluorescent 'timer' protein dsRed-E5, which has the unusual property of changing its fluorescence emission from green to red over ~ 16 h (ref. 3). The distribution of ANF-dsRed-E5-labelled LDCVs was similar to that of LDCVs labelled with ANF-EGFP (Fig. 1 and Supplementary Fig. 2). ANF-dsRed-E5 and ANF-EGFP co-localized with chromogranin A (CgA), an abundant protein found in the lumen of LDCVs, but not with a Golgi marker (Supplementary Fig. 2a–c). The mean number and size of ANF-dsRed-E5-labelled LDCVs (118 ± 19 , $n = 12$ vesicles per cell; diameter 380 nm, $n = 30$ vesicles from five cells) were not significantly different from those in cells labelled with ANF-EGFP (157 ± 17 vesicles per cell; $P = 0.15$). The mean number of ANF-dsRed-E5-labelled vesicles that were exclusively green or red was 27 ± 5 and 14 ± 5

per cell, respectively ($n = 12$ cells). LDCVs that possessed both red and green fluorescence appeared yellow (83 ± 13 per cell, $n = 12$).

Although all vesicles imaged in the present study were less than 72 h old—early in the 18-day lifespan of LDCVs⁷—even among these youthful vesicles there was spatial segregation according to age. In every cell possessing green LDCVs, the green vesicles were found predominantly at the cell periphery, whereas older vesicles (yellow and red) were found deeper in the same cell (Fig. 1a–c, e). The finding that the newest LDCVs had moved to the cell periphery, in preference to many pre-existing vesicles, argues against vesicle recruitment in order of assembly.

In close agreement with the ANF-EGFP data, nicotine stimulated the loss of $63 \pm 15\%$ of all dsRed-E5-labelled LDCVs, the overall number decreasing from 118 ± 12 to 43 ± 6 per cell (Fig. 1d; $n = 12$; $P = 0.001$). However, there was a striking pattern of secretion according to age: whereas $99 \pm 1\%$ of exclusively green (youngest) vesicles were released after maximal nicotine stimulation, $69 \pm 5\%$ of yellow vesicles were released, and none of the exclusively red vesicles were released (Fig. 1d, f). Thus, the youngest LDCVs are preferentially selected for secretion by nicotine. Because

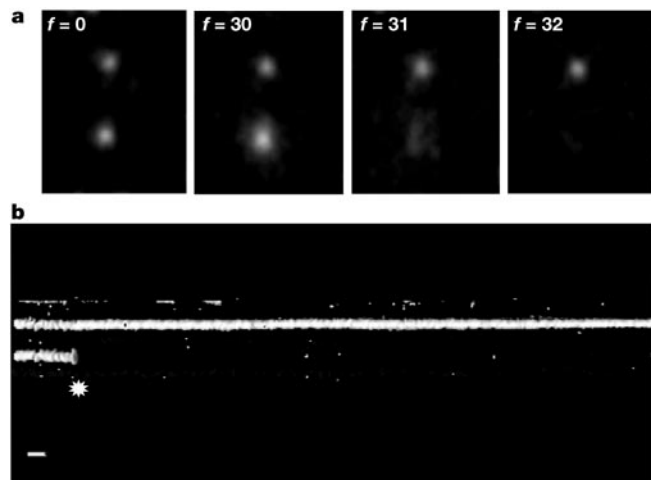


Figure 2 TIRFM imaging of transduced living BCCs resolves a membrane-proximal pool of labelled vesicles. **a**, Analysis of single-vesicle exocytosis. Single vesicles containing ANF-EGFP were seen to remain immobile in the x – y plane before exocytosis. An exocytotic flash was captured in a single 200-ms exposure and was seen to dissipate rapidly in the next image. Times are indicated as frame number, f , captured at 4.75 Hz. Stimulation was at $f = -43$. **b**, A 'two-dimensional + 1' reconstruction of this representative pair of vesicles. In **a** and **b** the upper vesicle remained immobile throughout the recording, whereas the lower vesicle underwent exocytosis from an immobile state. The asterisk in **b** indicates the flash ($f = 30$ in **a**) at which the lower vesicle underwent exocytosis. Time bar (bottom left in **b**), 21 frames, 4.2 s.

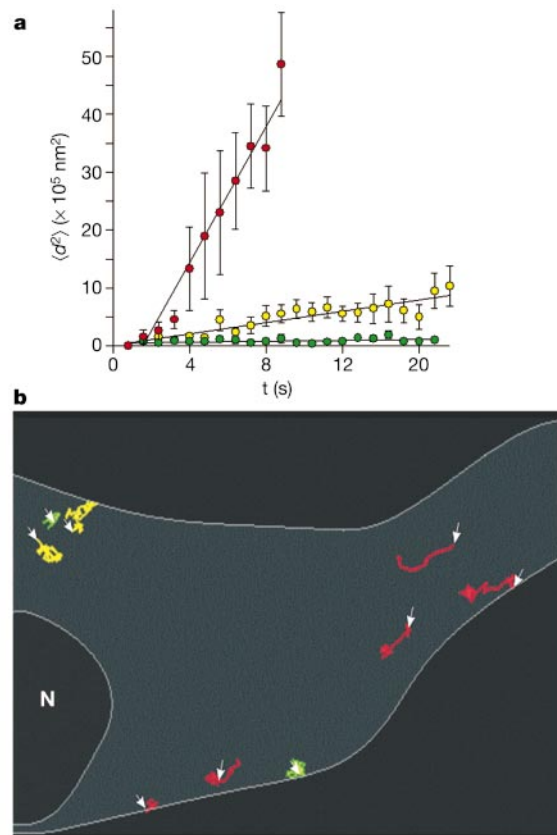


Figure 3 The mean squared displacement ($\langle d^2 \rangle$) was calculated for individual LDCVs from each ANF-dsRed-E5-coloured population, revealing differing mobilities for each group. **a**, Plots of $\langle d^2 \rangle$ against time for green (green circles), yellow (yellow circles) and red (red circles) LDCVs (189 data points, $n = 9$; 338 data points, $n = 13$; and 254 data points, $n = 25$ vesicles, respectively). Note that the data can be fitted according to the equation $\langle d^2 \rangle = 4D't$, as described previously^{11,12,17}. The apparent diffusion coefficient of green LDCVs is $D' = 7 \times 10^{-12} \text{ cm}^2 \text{ s}^{-1}$; that of yellow LDCVs is $D' = 3.7 \times 10^{-10} \text{ cm}^2 \text{ s}^{-1}$. Red LDCVs that exhibited rapid saltatory movements are fitted to a line with a calculated diffusion coefficient of $D' = 1.4 \times 10^{-9} \text{ cm}^2 \text{ s}^{-1}$. **b**, Composite image illustrating the directions of representative labelled granules from each group superimposed on a representation of a cell. Note that the green LDCVs are essentially immobile, yellow LDCVs exhibit a variety of motions and directions, and red LDCVs travel rapidly in straight lines. White arrows indicate the beginning of recording. N indicates the nucleus and the grey area represents the cytoplasm.

of the complex bleaching behaviour of dsRed-E5 and laser phototoxicity, cells were fixed either before or after stimulation, and images from each sample were compared.

Overexpression of ANF-EGFP had no significant effect on maximal catecholamine secretion or cellular catecholamine content (Fig. 1g). A comparison of transduced and non-transduced (control) cells revealed that the fractional release agreed with previously reported values for catecholamine^{8–10} and for endogenous ANF⁴. However, whereas in non-transduced cells the percentages of endogenous ANF and catecholamine released were similar (Fig. 1g), in transduced cells the proportion of ANF released was much larger: $85 \pm 18\%$ (Fig. 1g). This fractional release of ANF paralleled the large decrease in fluorescent vesicles in confocal studies (see above), indicating that the decrease in vesicle number is due to exocytosis (see also Fig. 2 for total internal reflection fluorescence microscopy (TIRFM) images of exocytosis) and that newly assembled LDCVs are preferentially recruited for exocytosis.

To confirm that the disappearance of newly assembled vesicles is due to exocytosis, we used TIRFM¹¹. During maximal stimulation with nicotine, exocytosis appeared as a short-lived flash of fluorescence generated as the ANF-EGFP escaped from the vesicles at the base of the cell (Fig. 2). An analysis of 15 single vesicle fusions from eight independent cells in which the EGFP-labelled LDCVs were clearly visible before fusion showed that peptide release originated from vesicles that had resided in a long-lived (>2 s) near-immobile state adjacent to the plasma membrane (Fig. 2b). Identical results were obtained with acridine-orange-loaded cells ($27 \pm 18\%$ of docked LDCV were released). Thus, newly assembled vesicles undergo exocytosis from an apparently membrane-anchored state, in agreement with previous studies in acridine-orange-loaded BCCs^{11,12}.

To examine functional differences between vesicle pools of distinct age we examined whether barium (Ba^{2+}), which reportedly acts on the reserve but not on the readily releasable pool in endocrine cells or neurons^{13,14}, could cause exocytosis of the newest (green) vesicles at the plasma membrane. In cell population assays 5 mM Ba^{2+} released a much larger fraction of the total catecholamine content ($60 \pm 9\%$ in control cells and $59 \pm 3\%$ in transduced cells; $n = 31$ cell preparations) than did nicotine (Fig. 1g), as has been previously reported¹⁵. However, the fraction of all labelled vesicles released in confocal images ($58 \pm 9\%$) was similar to that with nicotine stimulation ($63 \pm 15\%$, $P = 0.8$). In striking contrast to nicotine, Ba^{2+} failed to stimulate the release of green vesicles (27 ± 5 vesicles before stimulation, 30 ± 5 after stimulation), and released a greater fraction of yellow vesicles (83 ± 13 vesicles before stimulation, 14 ± 3 vesicles after stimulation, $n = 14$ cells, equivalent to $83 \pm 20\%$ of yellow vesicles) than did nicotine ($64 \pm 5\%$) (Fig. 1d–f). The mean number of red vesicles was also decreased by stimulation with Ba^{2+} ; however, experimental variability was higher for the oldest vesicle population, and the difference did not attain statistical significance.

The larger fractional release of catecholamine by Ba^{2+} compared with that by nicotine can be explained if Ba^{2+} , but not nicotine, stimulates a large fraction of unlabelled vesicles to undergo exocytosis. This indicates that red vesicles and unlabelled vesicles, both representing vesicle populations older than 16 h, are indistinguishable to the recruitment apparatus and that Ba^{2+} stimulation preferentially selects older vesicles for exocytosis. The amounts of catecholamine released by maximal stimulations with nicotine and Ba^{2+} were additive in population assays ($82 \pm 3\%$ release of total catecholamine), indicating that Ba^{2+} and nicotine might act by different mechanisms^{13,16}.

To confirm that Ba^{2+} cannot support the release of morphologically docked (that is, the youngest) LDCVs at the plasma membrane, we performed TIRFM imaging with acridine-orange-loaded cells stimulated maximally with nicotine or Ba^{2+} . In marked contrast to nicotine, which evoked the exocytosis of $27 \pm 18\%$ of

docked LDCVs ($n = 230$ vesicles from six cells), Ba^{2+} effected the release of no docked LDCVs ($n = 165$ vesicles from five cells), over the same time course of 4–10 min recording. Acridine orange stains all acidic vesicles in the cell, not just the newest vesicles. Although the age of LDCVs was undefined in these experiments with acridine orange, on the basis of our confocal imaging data we expect that a significant fraction of membrane-proximal LDCVs corresponded to the younger vesicle population. Thus, stimulation with Ba^{2+} , unlike that with nicotine, spares already-docked granules, which are most likely to represent the most recently assembled vesicles.

Analysis of ANF-dsRed-E5 labelled vesicles in living BCCs using confocal microscopy also revealed distinct mobilities for each coloured vesicle population. For individual granules, the mean squared displacement ($\langle d^2 \rangle$) was plotted as a function of observation time, and the initial slope of the plot was fitted with the equation $\langle d^2 \rangle = 4D't$, as described previously^{11,12,17} (Fig. 3). Markedly different values for the apparent diffusion constant (D') were found for the differently coloured LDCVs. Green vesicles had the lowest mobility, seeming to be in a near-immobile state near the membrane (Fig. 3), suggesting that they were morphologically docked¹². Yellow vesicles had a spread of mobilities, whereas the red LDCVs demonstrated two mobilities: either relatively immobile or rapidly mobile, travelling in straight lines in a saltatory manner (Fig. 3) with a mean velocity of $205 \pm 19 \text{ nm s}^{-1}$ and a mean peak velocity of $672 \pm 59 \text{ nm s}^{-1}$.

The use of a 'fluorescent timer' protein specifically targeted to LDCVs has allowed us to identify distinct pools of secretory vesicles, segregated according to age and distinguished by vesicle location, mobility and priority in recruitment for exocytosis. Newly assembled vesicles undergo rapid transport to the plasma membrane, where they are preferentially selected for exocytosis by nicotine, thus providing an explanation at the single-vesicle level for previous reports that newly synthesized peptides are quickly secreted^{18,19}. The rapid translocation of newly assembled LDCVs to the plasma membrane is in agreement with a recent study of vesicle dynamics in PC12 cells²⁰. LDCVs that do not undergo exocytosis are eventually (in less than 16 h) removed from the membrane, transported back into the cells and equilibrated with the large reserve vesicle pool, becoming indistinguishable from the pre-existing vesicles. These older vesicles are readily mobilized by Ba^{2+} , showing that vesicles of distinct age also differ functionally. Whether differential effects of nicotine and Ba^{2+} reflect differences in release mechanisms for vesicles of distinct ages^{13,14,16,21,22} remains to be elucidated.

Vesicle sorting according to age might provide a means whereby cells can regulate the nature of vesicles in the readily releasable pool. Preferential release of new vesicles would be advantageous if vesicular components (either membrane or cargo constituents) aged over time, perhaps becoming less functional²³. In addition, the preferential release of new vesicles would ensure that upregulation or downregulation of the genes encoding vesicle components is rapidly translated to changes in the readily releasable vesicle pool: new vesicles jump to the front of the queue ahead of older ones, possibly decreasing the time to effect from days (if vesicle recruitment followed rules of seniority) to only a few hours. □

Methods

Constructs and expression

The rat prepro-atrial natriuretic factor (ANF)-dsRed-E5 fusion construct was generated from an ANF-EGFP construct (a gift from E. S. Levitan, University of Pittsburgh (ref. 2)) ligated into the pTIMER Vector (Clontech). For expression in BCCs, these constructs were subcloned into either the pSFV1 (Gibco Life Sciences) or pSFV2gen²⁴ (a gift from K. Lundstrom, F. Hoffmann-La Roche) plasmids and Semliki Forest virus (SFV) generated essentially as described previously with similar transduction efficiencies 48 h after infection²⁵.

Confocal imaging and analysis

Cells on glass coverslips were imaged as described²⁵, but using simultaneous scanning with an excitation double dichroic beam splitter (488 and 568 nm fluorescein isothiocyanate/

tetramethylrhodamine isothiocyanate) using Leica TCS-NT and Zeiss LSM 510 confocal microscopes. Image data were deconvolved with Huygens II (Scientific Volume Imaging) and analysed with Bitplane software (Bitplane AG). The fluorescence emission of the ANF-ds-Red-E5 fusion protein in single LDCVs proceeded as described previously for dsRed-E5 (ref. 3). For vesicle mobility analysis, live BCCs were imaged at 0.8 Hz in HEPES-buffered DMEM supplemented with 10% (v/v) FCS. LDCV trajectories and velocities over time t were calculated with Metamorph (Universal Imaging Corp.). Time-lapse confocal laser scanning microscopy of exocytosis from ANF-dsRed-E5-labelled live cells proved impracticable for several technical reasons. First, the laser power required to illuminate the youngest (green) ds-Red-E5-labelled LDCVs was significantly higher (~10-fold) than that required to excite the EGFP-labelled LDCVs, presumably because of the lower quantum yield of immature dsRed than that of EGFP²⁶, and because of its slower maturation kinetics, which resulted in fewer fluorescent chromophores in younger vesicles. The high laser power required to discriminate both the location and 'colour' of individual ds-Red-E5-labelled vesicles was phototoxic over the time course (several minutes) required for analysis of secretion before and after stimulation. Second, bleaching of the dsRed-E5 chromophore was pronounced in live cells: red bleached faster than green, raising concerns about apparent colour changes due to diminished fluorescence resonance energy transfer between the maturing chromophores²⁶. For these reasons, we imaged cells that had been fixed either before or after secretion stimulation and mounted in Mowiol (Calbiochem).

Cell stimulation and biochemical analyses

Assays were performed on BCCs in the following buffer (concentrations in mM): 140 NaCl, 2.4 KCl, 2 CaCl₂, 1 MgCl₂, 10 HEPES pH 7.4 and 10 glucose. For all imaging and secretion assays, maximal stimulation was elicited by 20 μ M nicotine (or 5 mM Ba²⁺ as a replacement for the external Ca²⁺) as a secretagogue at 37 °C for 20 min. ANF was assayed in cell incubation media and cell lysates by radioimmunoassay with a rabbit antiserum against human ANF²⁷. Release and cellular content of catecholamine were assayed by fluorimetry in the presence of 0.6 M sodium acetate, 2.24 mM potassium ferricyanide, 3.35 mM sodium ascorbate (fluorescent emission determined at 517 nm after excitation at 410 nm).

TIRFM

A glass coverslip, with live cells attached to its upper face, was optically coupled to the upper face of an inverted trapezoidal prism with immersion oil (Zeiss; refractive index $n = 1.52$). Excitation at 488 nm from a Coherent Innova 90 argon-ion laser underwent total internal reflection at the glass ($n = 1.52$)/water ($n = 1.37$) interface at an incident angle of 61°, resulting in an exponentially decaying evanescent field with a space constant of approximately 110 nm. The fluorescence excited by the evanescent wave was imaged through an upright Zeiss Axioskop microscope equipped with a Zeiss $\times 63$ Achroplan (1.3 numerical aperture) water-immersion lens. An intensified charge-coupled device digital camera (intensified Pentamax; Princeton Instruments, Inc.) was used to capture TIRFM images at 200 ms per frame with a 10-ms recycle time. Pixel size was 15 $\times$ 15 μ m, equivalent to ~177.5 $\times$ 177.5 nm in the specimen plane. Image capture and analysis were performed with MetaMorph. A single-vesicle fusion was apparent as a short-lived increase in fluorescence spot intensity and disappearance of the spot in a rapidly dispersing cloud. Thresholded fluorescent pixels within the time series were stacked and surface rendered with Isosurface (Bitplane), to provide a 'two-dimensional + 1' time-stack projection. Thus, an immobile vesicle appears as a solid bar and an exocytotic flash as a 'mushroom cap' (Fig. 2). Some exocytotic flashes were observed in which the vesicle could not be detected before fusion. These flashes were more diffuse and occurred in isolation from any visible docked vesicles, suggesting that they arose from exocytosis at membrane that had buckled up beyond the reach of evanescent wave excitation. Such events were not analysed.

Data analysis

Data are presented as means $\pm$ s.e.m.

Received 20 August; accepted 13 December 2002; doi:10.1038/nature01389.

1. Morgan, A. & Burgoyne, R. D. in *Seminars in Cell and Developmental Biology* (eds Burgoyne, R. D. & Mombaerts, P.) 141–149 (Academic, Cambridge, 1997).
2. Burke, N. V. *et al.* Neuronal peptide release is limited by secretory granule mobility. *Neuron* **19**, 1095–1102 (1997).
3. Terskikh, A. *et al.* 'Fluorescent timer': protein that changes colour with time. *Science* **290**, 1585–1588 (2000).
4. Nguyen, T. T., Ong, H. & De Lean, A. Secretion and biosynthesis of atrial natriuretic factor by cultured adrenal chromaffin cells. *FEBS Lett.* **231**, 393–396 (1988).
5. Cuchillo-Ibanez, I., Michelena, P., Albillos, A. & Garcia, A. G. A preferential pole for exocytosis in cultured chromaffin cells revealed by confocal microscopy. *FEBS Lett.* **459**, 22–26 (1999).
6. Phillips, J. H. Dynamic aspects of chromaffin granule structure. *Neuroscience* **7**, 1595–1609 (1982).
7. Corcoran, J. J., Wilson, S. P. & Kirshner, N. Flux of catecholamines through chromaffin vesicles in cultured bovine adrenal medullary cells. *J. Biol. Chem.* **259**, 6208–6214 (1984).
8. Burgoyne, R. D. Control of exocytosis in adrenal chromaffin cells. *Biochim. Biophys. Acta* **1071**, 174–202 (1991).
9. Neher, E. & Zucker, R. S. Multiple calcium-dependent processes related to secretion in bovine chromaffin cells. *Neuron* **10**, 21–30 (1993).
10. Parsons, T. D., Coorsen, J. R., Horstmann, H. & Almers, W. Docked granules, the exocytic burst, and the need for ATP hydrolysis in endocrine cells. *Neuron* **15**, 1085–1096 (1995).
11. Steyer, J. A., Horstmann, H. & Almers, W. Transport, docking and exocytosis of single secretory granules in live chromaffin cells. *Nature* **388**, 474–478 (1997).
12. Oheim, M., Loeke, D., Stuhmer, W. & Chow, R. H. The last few milliseconds in the life of a secretory granule. Docking, dynamics and fusion visualized by total internal reflection fluorescence

- microscopy (TIRFM). *Eur. Biophys. J.* **27**, 83–98 (1998).
13. Seward, E. P., Chernetskaya, N. I. & Nowycky, M. C. Ba²⁺ ions evoke two kinetically distinct patterns of exocytosis in chromaffin cells, but not in neurohypophyseal nerve terminals. *J. Neurosci.* **16**, 1370–1379 (1996).
14. Neves, G., Neef, A. & Lagnado, L. The actions of barium and strontium on exocytosis and endocytosis in the synaptic terminal of goldfish bipolar cells. *J. Physiol. (Lond.)* **535**, 809–824 (2001).
15. TerBush, D. R. & Holz, R. W. Barium and calcium stimulate secretion from digitonin-permeabilized bovine adrenal chromaffin cells by similar pathways. *J. Neurochem.* **58**, 680–687 (1992).
16. Heldman, E., Levine, M., Raveh, L. & Pollard, H. B. Barium ions enter chromaffin cells via voltage-dependent calcium channels and induce secretion by a mechanism independent of calcium. *J. Biol. Chem.* **264**, 7914–7920 (1989).
17. Han, W., Ng, Y. K., Axelrod, D. & Levitan, E. S. Neuropeptide release by efficient recruitment of diffusing cytoplasmic secretory vesicles. *Proc. Natl Acad. Sci. USA* **96**, 14577–14582 (1999).
18. Ogawa, T., Vatta, M., Bruneau, B. G. & de Bold, A. J. Characterization of natriuretic peptide production by adult heart atria. *Am. J. Physiol.* **276**, H1977–H1986 (1999).
19. Chen, T. T. *et al.* Relative importance of newly synthesized and stored hormone to basal secretion by growth hormone and prolactin cells. *Endocrinology* **125**, 1904–1909 (1989).
20. Rudolf, R., Salm, T., Rustom, A. & Gerdes, H. H. Dynamics of immature secretory granules: role of cytoskeletal elements during transport, cortical restriction, and F-actin-dependent tethering. *Mol. Biol. Cell* **12**, 1353–1365 (2001).
21. Sugita, S. *et al.* Synaptotagmins form a hierarchy of exocytotic Ca²⁺ sensors with distinct Ca²⁺ affinities. *EMBO J.* **21**, 270–280 (2002).
22. Li, C. *et al.* Ca²⁺-dependent and Ca²⁺-independent activities of neural and nonneural synaptotagmins. *Nature* **375**, 594–599 (1995).
23. Winkler, H. & Fischer-Colbrie, R. Regulation of the biosynthesis of large dense-core vesicles in chromaffin cells and neurons. *Cell. Mol. Neurobiol.* **18**, 193–209 (1998).
24. Lundstrom, K. *et al.* Novel mutant Semliki Forest virus vectors: gene expression and localization studies in neuronal cells. *Histochem. Cell. Biol.* **115**, 83–91 (2001).
25. Duncan, R. R. *et al.* High-efficiency Semliki Forest virus-mediated transduction in bovine adrenal chromaffin cells. *Biochem. J.* **342**, 497–501 (1999).
26. Baird, G. S., Zacharias, D. A. & Tsien, R. Y. Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral. *Proc. Natl Acad. Sci. USA* **97**, 11984–11989 (2000).
27. Freestone, S. *et al.* The renal effects of atrial natriuretic peptide in man are not attenuated by (+)-sulpiride. *Br. J. Clin. Pharmacol.* **27**, 13–18 (1989).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank A. Morgan for the catecholamine assay protocol; J. Pryde for the polyclonal anti-Golgi serum; D. J. Webb and N. Johnston for the anti-ANF serum and RIA protocol; and L. Sharp for assistance with confocal microscopy. This work was funded by Wellcome Trust grants and an NIH grant to D.K.A., R.H.C. and M.J.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.H.C. (e-mail: rchow@hsc.usc.edu) or M.J.S. (e-mail: mike.shipston@ed.ac.uk).

Functional analysis of an archaeobacterial voltage-dependent K⁺ channel

Vanessa Ruta, Youxing Jiang, Alice Lee, Jiayun Chen & Roderick MacKinnon

Howard Hughes Medical Institute, Laboratory of Molecular Neurobiology and Biophysics, Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

All living organisms use ion channels to regulate the transport of ions across cellular membranes¹. Certain ion channels are classed as voltage-dependent because they have a voltage-sensing structure that induces their pores to open in response to changes in the cell membrane voltage. Until recently, the voltage-dependent K⁺, Ca²⁺ and Na⁺ channels were regarded as a unique development of eukaryotic cells, adapted to accomplish specialized electrical signalling, as exemplified in neurons. Here we present the functional characterization of a voltage-dependent K⁺ (K_V) channel from a hyperthermophilic archaeobacterium from an

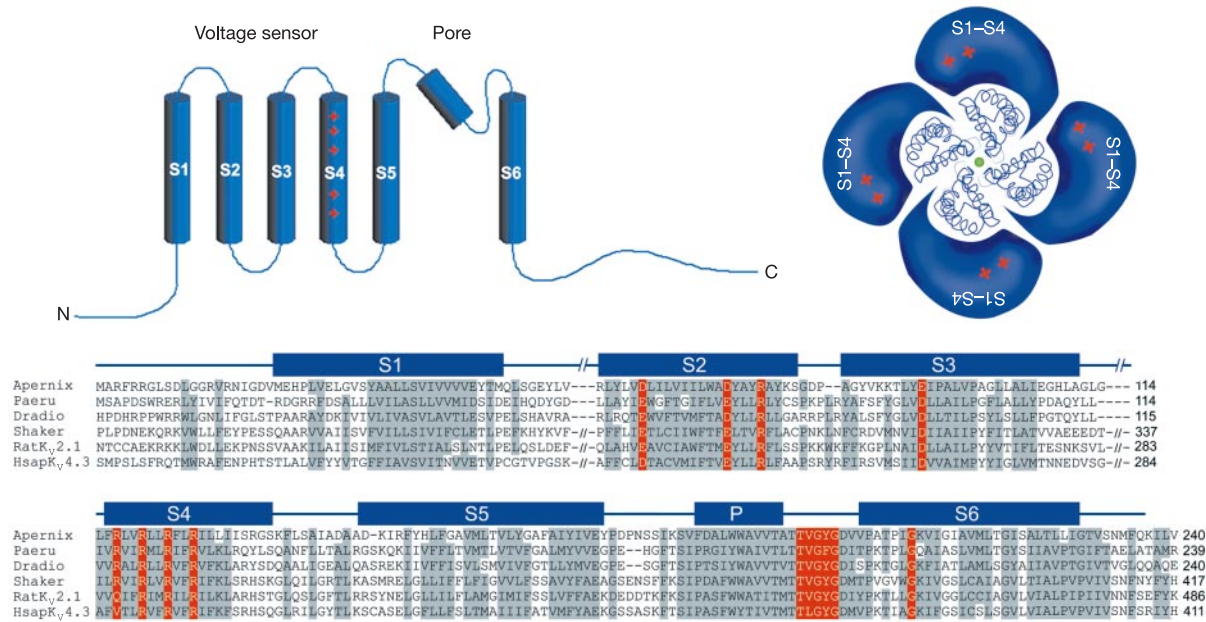


Figure 1 Architecture of a voltage-dependent K⁺ channel. Transmembrane-spanning segments (S1–S6) are labelled (blue segments); four subunits surround the pore. S1–S4 form the voltage sensor and S5–S6 form the pore, represented by the KcsA K⁺ channel structure (backbone model). In sequences of prokaryotic and eukaryotic K_V channels, regions of high homology are coloured in grey; functionally important residues are

coloured red. Alignment was made with ClustalW followed by manual adjustment and exclusion of loops. The K⁺ channels are: Apernix, *Aeropyrum pernix* (GI: 5104624); Paeru, *Pseudomonas aeruginosa* (GI: 15596693); Dradio, *Deinococcus radiodurans* (GI: 15805856); Shaker, *Drosophila melanogaster* (GI: 13432103); RatK_V2.1, *Rattus norvegicus* (GI: 24418849); HsapK_V4.3, *Homo sapiens* (GI: 5059060).

oceanic thermal vent. This channel possesses all the functional attributes of classical neuronal K_V channels. The conservation of function reflects structural conservation in the voltage sensor as revealed by specific, high-affinity interactions with tarantula venom toxins, which evolved to inhibit eukaryotic K_V channels.

The sequencing of prokaryotic genomes has revealed genes that seem to encode voltage-dependent ion channels. One such gene, a Na⁺ selective channel, was shown to exhibit voltage-dependent gating when expressed in a mammalian cell line². Figure 1 shows three prokaryotic genes (top three sequences) that are closely related

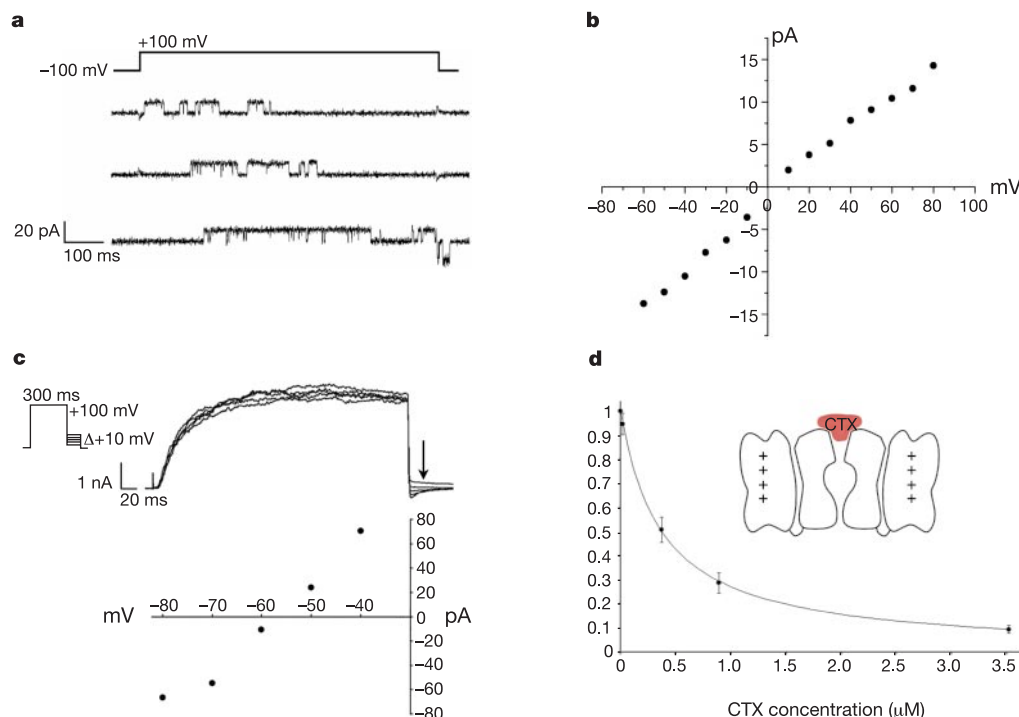


Figure 2 Properties of the ion pathway. **a**, Single-channel traces were recorded at +100 mV in 150 mM KCl, 10 mM HEPES, pH 7.0. **b**, Single-channel current as a function of membrane voltage in above solutions. **c**, Determination of K⁺ selectivity. Currents were measured at various test potentials (arrow) after depolarizing to +100 mV in solutions of 150 mM KCl, 10 mM HEPES, pH 7.0 inside and 15 mM KCl, 135 mM NaCl, 10 mM

HEPES, pH 7.0 outside. Leak and capacitive currents were subtracted (see Methods). The reversal potential is 56.5 ± 1.2 mV ($n = 4$). **d**, CTX recognizes a conserved pore structure, fitting like a key in a lock (inset). Fraction of unblocked current (I/I_{max}) is plotted (mean $\pm$ s.e.m., $n = 3$) as a function of CTX concentration and fit to $I/I_{max} = (1 + [CTX]/K_D)^{-1}$ with a K_D of 0.4 μ M.

to eukaryotic K_V channels (bottom three sequences). K_V channels are thought to contain six membrane-spanning segments per subunit (S1–S6): S1–S4 comprise the voltage sensor, whereas S5–S6 comprise the ion-selective pore. Four identical subunits encircle a central ion pathway. Functional studies of eukaryotic K_V channels have identified a number of important amino acids (Fig. 1, residues highlighted in red). ‘Signature sequence’ amino acids between P and S6 form the selectivity filter, conserved inside the pore of all known K^+ channels³, and a glycine residue in S6 serves as a hinge, allowing the pore to open⁴. Four arginine residues at every third position in S4 and additional charged residues in S2 and S3 underlie voltage-dependent gating. It is thought that the S4 arginine residues move through the membrane’s electric field to open the pore^{5,6}. Conservation of the residues in red (Fig. 1) implies

that the prokaryotic channels may function very much like eukaryotic K_V channels with respect to both K^+ selective conduction and voltage-dependent gating. If this is indeed the case then prokaryotic K_V channels provide an inroad towards understanding the structural basis of the voltage-dependent gating mechanism.

Are these prokaryotic K^+ channels truly gated by voltage? Until now, essentially nothing was known about their functional properties because they cannot easily be studied with the expression systems used for eukaryotic channels. We present here the functional characterization of a prokaryotic K_V channel (K_VAP) cloned from *Aeropyrum pernix* (Apernix, Fig. 1); a hyperthermophilic archaeon isolated from an oceanic thermal vent off the coast of Japan (GI: 5104624). Although the physiological role of this channel is unknown, we show that it is functionally and structurally very similar to eukaryotic K_V channels.

K_VAP channels were expressed in *Escherichia coli*, extracted with decylmaltoside, purified and reconstituted into planar lipid bilayers of 1-palmitoyl-2-oleoyl-phosphatidylglycerol (POPG) and 1-palmitoyl-2-oleoyl-phosphatidylethanolamine (POPE) for functional studies. By controlling the protein to lipid ratio used in the reconstitution, we were able to study both single-channel and macroscopic current records. K_VAP channels have a large conductance—the slope of the single-channel current–voltage relationship recorded in solutions containing 150 mM KCl and 10 mM HEPES, pH 7.0, on both sides of the membrane shows a conductance of approximately 170 pS (Fig. 2a, b). The presence of the K^+ channel signature sequence indicates that the K_VAP pore should be strongly selective for K^+ versus Na^+ ions. To examine ion selectivity we measured the reversal potential of macroscopic tail currents in a tenfold K^+ gradient by substituting 135 mM NaCl for 135 mM KCl in the solution on one side of the membrane (Fig. 2c). The measured reversal potential is -56.5 ± 1.2 mV, which is near the Nernst potential for a perfectly K^+ selective pore at room temperature (21 °C).

How structurally similar is the pore of the K_VAP channel to eukaryotic K^+ channel pores? To answer this question we looked for the ability of a small protein toxin from scorpion venom to inhibit the K_VAP channel. Venomous animals, such as scorpions, exploit the conservation of ion-channel structure by producing a toxin that recognizes a structural feature common to an entire family of ion channels⁷. By making many sequence variants of the same basic toxin structure, a scorpion can inhibit virtually every member of an ion channel family. The scorpion *Leiurus quinquestriatus hebraeus* specializes in a family of pore-blocking toxins, exemplified by charybdotoxin (CTX), which fit, like a key to a lock, to the pore entryway of K^+ channels (Fig. 2d)^{8,9}. We find that CTX inhibits the K_VAP channel with a dissociation constant (K_d) of about 0.4 μ M. Although the affinity is not very high, single point mutations in the pore of eukaryotic K^+ channels can reduce toxin affinity by greater than 1,000-fold^{10,11}. We emphasize that CTX would not bind to the K_VAP channel if its pore were not very similar in structure to that of eukaryotic K^+ channels. Given the high degree of amino acid sequence conservation within the pore, near identity in the three-dimensional structure of the K_VAP channel and the *Shaker* K^+ channel, for example, is expected (Fig. 1).

We next investigated whether gating in the K_VAP channel is voltage-dependent. Figure 3 indicates that this is indeed the case, with qualitative and quantitative properties that are similar to eukaryotic K_V channels. To determine the orientation of channels incorporated into planar lipid bilayers, we used CTX, which causes inhibition by binding only to the extracellular side. We find that K_VAP channels open in response to membrane depolarization; that is, when the voltage of the CTX-insensitive (intracellular) side of the membrane is made positive relative to the CTX-sensitive (extracellular) side. Figure 3a shows a representative family of current traces elicited by voltage steps to increasingly positive values from a holding membrane voltage of -100 mV. From such records

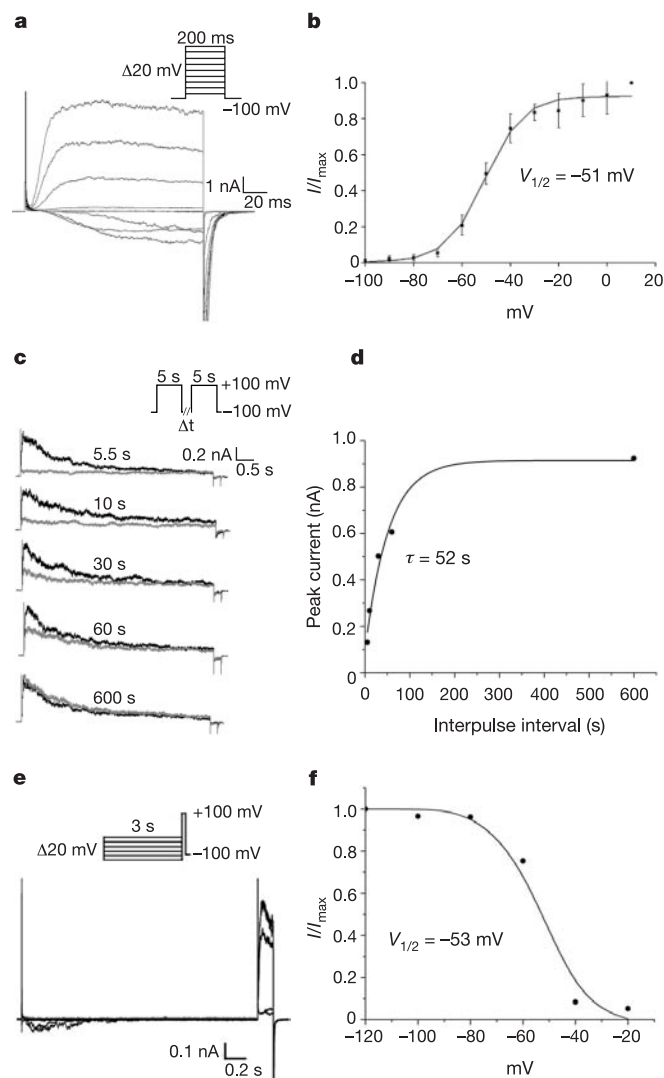


Figure 3 Voltage-dependent gating properties. **a**, Channel activation. Currents were recorded during the voltage protocol shown (inset) in the presence of 150 mM KCl. **b**, The voltage activation curve. The fraction of maximal current (I/I_{max} , mean $\pm$ s.e.m., $n = 6$) is graphed as a function of test potential (see Methods). The smooth curve corresponds to equation (1) with $V_{1/2}$ and Z of -51 mV and 3.2 , respectively. **c**, Channel inactivation. Two 5-s pulses (black first, grey second) to $+100$ mV are separated by holding at -100 mV for the indicated times. **d**, Fraction of recovery as a function of interpulse duration for currents shown in **c** is fitted to a single exponential with time constant 52 s. **e**, Steady-state inactivation. Currents were recorded during the voltage protocol shown (inset) in 150 mM KCl. **f**, Fraction of activatable currents measured in **e** is graphed as a function of holding voltage. The smooth curve is a fit to a two-state Boltzmann equation.

we evaluated the voltage dependence of channel opening (Fig. 3b). Data were fitted with the Boltzmann equation:

$$\frac{I}{I_{\max}} = \frac{1}{1 + \exp\left(-\frac{ZF}{RT}(V - V_{1/2})\right)} \quad (1)$$

where F is the Faraday constant, $I/I_{\max}$ is the fraction of the maximal current, $V_{1/2}$ is the voltage at which the channels have reached 50% of their maximum fraction open, equal to -51 mV, R is the gas constant, T is temperature and Z is the apparent valence of voltage dependence, equal to 3.2. We note that the total protein charge moving through the membrane electric field, known as the gating charge, is 12–14 electron charge equivalents in the *Shaker* K^+ channel^{12–14}. But the gating charge should not be confused with Z in the voltage dependence of channel opening, which is around 3.0 for the *Shaker* K^+ channel¹⁵. K_V AP channels are strongly voltage-dependent, opening as a function of membrane voltage very much like *Shaker* and other neuronal K_V channels.

Even many subtle aspects of the voltage-dependent gating of K_V AP are similar to that observed in neuronal K_V channels. For example, the sigmoidal time dependence of opening immediately after depolarization in Fig. 3a, thought to reflect multiple transitions along the reaction pathway to opening, looks similar to *Shaker* K^+ channels and K_V channels in squid axons that underlie the Hodgkin–Huxley theory of the action potential^{16–18}. The main obvious distinction is the slowness of the overall rates at which K_V AP channels open and close. But K_V AP is derived from an organism that thrives at temperatures near 95°C , whereas we are studying the channel's properties at room temperature.

Another aspect of gating in many eukaryotic K_V channels is a

phenomenon called inactivation, referring to spontaneous pore closure during sustained membrane depolarization. K_V AP channels also exhibit inactivation, as evidenced by the gradual reduction in the current (50% decay in about 800 ms at $+100$ mV) during long depolarizing steps (Fig. 3c), and slow recovery from inactivation after returning the membrane to its holding voltage (50% in 35 s, 90% in 120 s; Fig. 3c, d). The fraction of channels that become inactivated as a function of the membrane voltage follows a Boltzmann equation (Fig. 3e, f). This voltage dependence of 'steady-state inactivation' is similar to that observed in many eukaryotic K_V channels.

In eukaryotic K_V channels two varieties of inactivation have been described: N type, in which the cytoplasmic amino terminus occludes the pore after the voltage-dependent gate opens¹⁹, and C type, which is less well understood²⁰. We do not yet know which mechanism is operating in K_V AP, but we suspect that it is the C type. The important point here is that the voltage-dependent activation and inactivation in the K_V AP channel is phenomenologically very similar to what has been observed in eukaryotic K_V channels studied by electrophysiologists for decades.

Do the voltage-dependent gating properties of the K_V AP channel arise from a structurally conserved voltage sensor that is shared by prokaryotic and eukaryotic K_V channels? The amino acid sequence conservation indicates, unequivocally, that the answer is yes (Fig. 1). But just how conserved is it? If the voltage sensor is conserved to the extent of the K^+ selective pore, then natural toxins intended to inhibit eukaryotic voltage sensors might cross-react with K_V AP, just as CTX cross-reacts with the pore. We tested this possibility using venom from the Chilean rose tarantula, *Grammostola spatulata*.

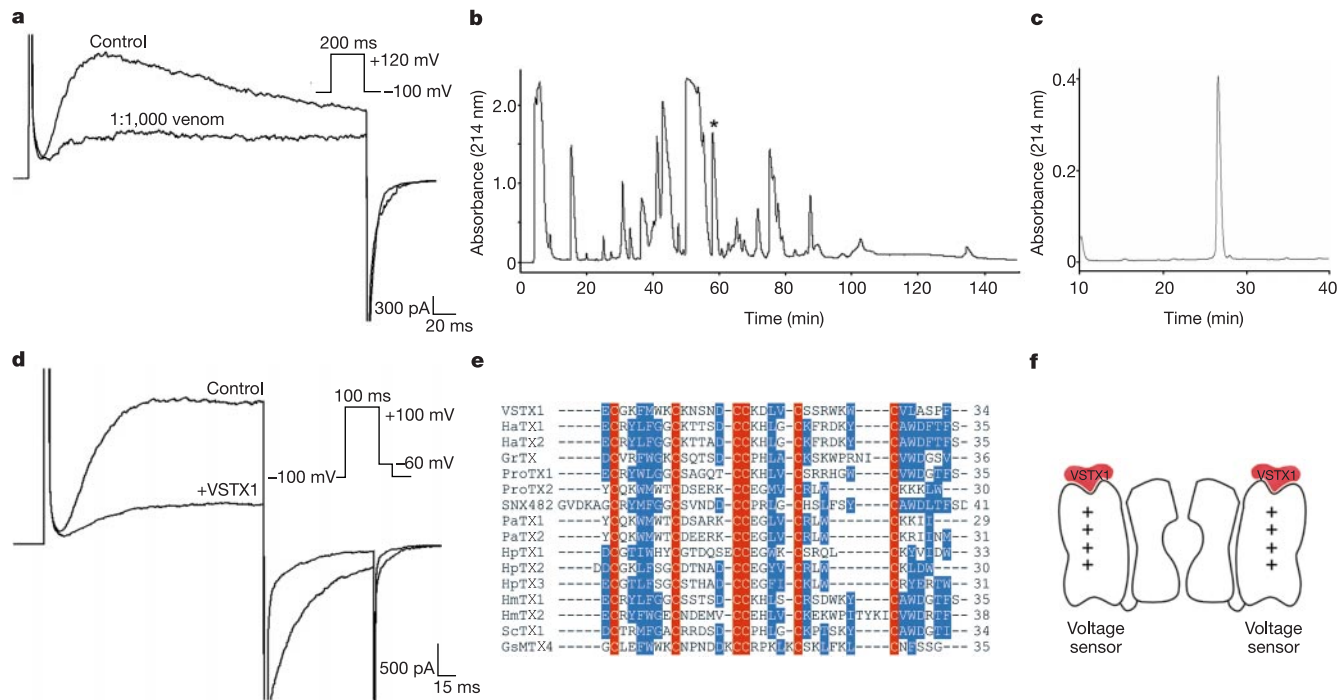


Figure 4 Inhibition by *G. spatulata* toxins. **a**, A 1:1,000 dilution of spider venom inhibits currents elicited by depolarization to $+120$ mV. **b**, Reverse-phase HPLC chromatogram of venom showing the fraction with VSTX1. **c**, VSTX1 is purified to homogeneity by a second HPLC run (see Methods). **d**, Effect of about 30 nM purified VSTX1 on the K_V AP channel, 70 min after toxin addition. **e**, Sequence analysis of VSTX1 and other spider gating-modifier toxins. Highly conserved residues are coloured in blue and the six cysteine residues that determine the tarantula toxin fold are highlighted in red. The spider toxins are: VSTX1, voltage-sensor toxin 1; HaTX1 and HaTX2, hanatoxins (GI: 1245765

and GI: 1245766, respectively); GrTX, grammatxin (GI: 451235); ProTX1 (GI: 25091452); ProTX2 (GI: 25091451); SNX482 (GI: 7388243); PaTX1 and PaTX2, phrixotoxins²⁹; HpTX1, HpTX2 and HpTX3, heteropodatoxins (GI: 17433177, GI: 17433178 and GI: 17433179, respectively); HmTX1 and HmTX2, heteroscodratoxins³⁰; ScTX1, stomatocin 1 (ref. 30); GsMTx4, *Grammostola* mechanotoxin number 4 (GI: 25412346). **f**, Diagram of VSTX1 recognizing a conserved structure of the voltage sensor.

This venom contains toxins such as hanatoxin (HaTX) and other structurally related proteins that inhibit eukaryotic voltage-dependent ion channels by binding to their voltage sensors^{21–23}. On addition of a 1:1,000 dilution of the whole venom to the extracellular side of the membrane, we observed inhibition of K_VAP currents elicited by a +120-mV depolarizing step (Fig. 4a). Next, to isolate and identify active components, we fractionated the venom by reversed-phase high-performance liquid chromatography (HPLC) and tested individual peaks for activity. Several toxins of different masses reproduced the inhibitory effect of the total venom. We focused on one of these, which we call voltage-sensor toxin 1 (VSTX1). VSTX1 can be purified to homogeneity by two successive HPLC gradients (Fig. 4b, c). This toxin inhibits the K_VAP channel with very slow apparent binding kinetics, a characteristic of other gating-modifier toxins^{21,23}. The slow kinetics makes it difficult to reach equilibrium with sub-saturating concentrations before the membrane breaks, but more than half inhibition was achieved after 70 min of periodic membrane depolarizing steps in the presence of 30 nM VSTX1 (Fig. 4d). Therefore, VSTX1 inhibits K_VAP with at least low nanomolar affinity.

The amino acid sequence of VSTX1 was determined through a combination of N-terminal sequence analysis and matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF) mass spectrometry. VSTX1 is related to HaTX and other gating-modifier toxins from tarantula venoms (Fig. 4e). The most conserved element of these toxins is the six cysteine residues that form three disulphide bonds and thus define their rigid three-dimensional structure, a structure that has evolved to recognize and bind to the voltage sensor. Of note, some of these toxins promiscuously react with the voltage sensors of eukaryotic voltage-dependent K⁺, Ca²⁺ and Na⁺ channels^{24,25}, underscoring the structural conservation of the voltage sensor in channels whose pores have diverged to allow different cation selectivity. It is clear that the K_VAP channel, with its sensitivity to tarantula toxins, will be a basic structural model for all members of the voltage-dependent cation channel family (K⁺, Na⁺ and Ca²⁺ channels).

We have shown that the K_VAP channel is very similar to eukaryotic K_V channels in both its functional and structural properties. The voltage dependence of K_VAP channel opening is essentially the same as in eukaryotic K_V channels. The evidence that K_VAP is structurally the same as its eukaryotic counterparts is based on amino acid sequence conservation and molecular recognition by inhibitors of both the pore and the voltage sensor. There is no evolutionary pressure for a tarantula native to the dry scrublands of Chile or a scorpion from the deserts of the Middle East to recognize receptor sites on a K_V channel from an archaeobacterium that lives near an oceanic thermal vent off the coast of Japan. The specific protein–protein interactions described here, which transcend geographical and traditional evolutionary boundaries, highlight the existence of highly conserved molecular structures underlying both ion selectivity and voltage-dependent gating in K_V channels throughout nature. □

Methods

Protein preparation and analysis

A sample of *Aeropyrum pernix* was obtained from the Japan Collection of Microorganisms. *Aeropyrum pernix* cultures were grown in a solution of autoclaved sea water supplemented with bacto yeast extract, trypticase peptone and sodium thiosulphate for three days in an oil bath maintained at 95 °C. *Aeropyrum pernix* genomic DNA was collected by standard procedures. The gene coding for K_VAP starting from methionine 14 was cloned by polymerase chain reaction (PCR) amplification of the genomic DNA and inserted into the protein expression vector pQE60 (Qiagen) between *Nco*I and *Bgl*II restriction endonuclease sites with a thrombin cleavage site between a carboxy-terminal hexahistidine sequence and the channel. Channel protein was expressed in XL1-blue cell cultures grown in LB medium supplemented with 10 mM BaCl₂ on induction with 0.4 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Expressed protein was extracted with 40 mM decylmaltoside (DM) and purified on a Talon Co²⁺ affinity column (Clontech). The protein was maintained in 5 mM DM, 20 mM Tris, pH 8.0, and 100 mM KCl. Nonspecifically bound protein was washed using 15 mM imidazole added to the

above buffer, and the channel then eluted with 400 mM imidazole. Immediately after elution, 1.0 unit of thrombin (Roche) per 3.0 mg channel was added to cleave the hexahistidine sequence overnight at room temperature. Protein was concentrated to about 15 mg ml^{−1} and run on a Superdex-200 (10/30) column (Pharmacia) in the above buffer. MALDI–TOF mass spectrometry (PerSeptive Biosystems Voyager-STR) and N-terminal sequencing analysis (Rockefeller University Protein/DNA Technology Center) indicated that the K_VAP protein undergoes a modification during expression in which the first five residues of the encoded construct are replaced with a single leucine residue in the expressed channel protein.

Electrophysiology

K_VAP channels were reconstituted from DM into lipid vesicles as described²⁶, to give a final protein concentration of between 20 to 200 μg ml^{−1}. Planar lipid bilayers of POPE (15 mg ml^{−1}) and POPG (5 mg ml^{−1}) in decane were painted over a 300-μm hole in a polystyrene partition separating the internal and external solutions. To induce fusion of channel-containing vesicles, solution on the side to which vesicles were added (*cis*) contained 150 mM KCl, 10 mM HEPES, pH 7.0 and the opposite side (*trans*) contained 15 mM KCl, 10 mM HEPES, pH 7.0. After the appearance of channels in the membrane, ion concentration on the *trans* side was increased to 150 mM KCl or 15 mM KCl plus 135 mM NaCl. Comparison of Figs 2c and 3a show that the ionic composition influences activation kinetics. Leak and capacitive currents were subtracted for Fig. 2c by inactivating K⁺ channels with a rapid pulse sequence, recording a leak template and subtracting it from traces with active channels. For voltage activation curves (Fig. 3b) the membrane was stepped to various test voltages and then returned to −100 mV. Currents were recorded 10 ms after the return to −100 mV, normalized to the current measured following the +10-mV test voltage, and graphed as a function of test voltage. Membrane voltage was controlled and current recorded using an Axopatch 200B amplifier with a Digidata 1322A analogue to digital converter and Axoclamp software (Axon Instruments).

Toxin biochemistry

CTX was produced recombinantly as previously described²⁷. Venom from *G. spatulata* spiders was purchased from Spider Pharm. Venom was diluted tenfold in 0.1% trifluoroacetic acid (TFA) in water, clarified by centrifugation and filtered through a 0.22-μm cellulose acetate filter. The resulting supernatant was fractionated in two steps using an Agilent 1100 Series high-performance liquid chromatographer (HPLC) with a C-18 reverse-phase, 5-μm, 80 Å column. The initial fractionation of venom used a linear gradient at a flow rate of 1 ml min^{−1} from 22% to 56% mobile phase B over 150 min, where mobile phase A was 0.1% TFA in water and mobile phase B was 10% water, 90% acetonitrile and 0.1% TFA. The peak eluting at about 59 min contained mostly VSTX1. Homogeneous VSTX1 could be obtained through a second purification step using a 2-min isocratic flow of 30% mobile phase B followed by a linear gradient at a flow rate of 1 ml min^{−1} from 30% to 42% mobile phase B over 40 min. Purified VSTX1 was analysed by MALDI–TOF mass spectrometry (PerSeptive Biosystems Voyager-STR). The first 32 residues of the VSTX1 sequence were determined using Edman chemistry (Rockefeller University Protein/DNA Technology Center). The identity and order of the final two residues was determined using tandem mass spectrometry on the C-terminal toxin fragment obtained from a Lys-C proteolysis of reduced and alkylated VSTX1. VSTX1 concentrations were determined by 280-nm absorbance using the calculated extinction coefficient²⁸.

Received 16 December 2002; accepted 7 February 2003; doi:10.1038/nature01473.

Published online 2 March 2003.

- Hille, B. *Ion Channels of Excitable Membranes* (Sinauer, Sunderland, Massachusetts, 2001).
- Ren, D. *et al.* A prokaryotic voltage-gated sodium channel. *Science* **294**, 2372–2375 (2001).
- Heginbotham, L., Lu, Z., Abramson, T. & MacKinnon, R. Mutations in the K⁺ channel signature sequence. *Biophys. J.* **66**, 1061–1067 (1994).
- Jiang, Y. *et al.* The open pore conformation of potassium channels. *Nature* **417**, 523–526 (2002).
- Sigworth, F. J. Voltage gating of ion channels. *Q. Rev. Biophys.* **27**, 1–40 (1994).
- Bezannila, F. The voltage sensor in voltage-dependent ion channels. *Physiol. Rev.* **80**, 555–592 (2000).
- MacKinnon, R., Cohen, S. L., Kuo, A., Lee, A. & Chait, B. T. Structural conservation in prokaryotic and eukaryotic potassium channels. *Science* **280**, 106–109 (1998).
- Miller, C., Moczydlowski, E., Latorre, R. & Phillips, M. Charybdotoxin, a protein inhibitor of single Ca²⁺-activated K⁺ channels from mammalian skeletal muscle. *Nature* **313**, 316–318 (1985).
- MacKinnon, R. & Miller, C. Mechanism of charybdotoxin block of the high-conductance, Ca²⁺-activated K⁺ channel. *J. Gen. Physiol.* **91**, 335–349 (1988).
- Garcia, M. L., Garcia-Calvo, M., Hidalgo, P., Lee, A. & MacKinnon, R. Purification and characterization of three inhibitors of voltage-dependent K⁺ channels from *Leiurus quinquestratus* var. *hebraeus* venom. *Biochemistry* **33**, 6834–6839 (1994).
- Goldstein, S. A., Pheasant, D. J. & Miller, C. The charybdotoxin receptor of a Shaker K⁺ channel: peptide and channel residues mediating molecular recognition. *Neuron* **12**, 1377–1388 (1994).
- Schoppa, N. E., McCormack, K., Tanouye, M. A. & Sigworth, F. J. The size of gating charge in wild-type and mutant Shaker potassium channels. *Science* **255**, 1712–1715 (1992).
- Aggarwal, S. K. & MacKinnon, R. Contribution of the S4 segment to gating charge in the Shaker K⁺ channel. *Neuron* **16**, 1169–1177 (1996).
- Seoh, S. A., Sigg, D., Papazian, D. M. & Bezannila, F. Voltage-sensing residues in the S2 and S4 segments of the Shaker K⁺ channel. *Neuron* **16**, 1159–1167 (1996).
- Yifrach, O. & MacKinnon, R. Energetics of pore opening in a voltage-gated K(+) channel. *Cell* **111**, 231–239 (2002).
- Zagotta, W. N., Hoshi, T. & Aldrich, R. W. Shaker potassium channel gating. III: Evaluation of kinetic models for activation. *J. Gen. Physiol.* **103**, 321–362 (1994).
- Schoppa, N. E. & Sigworth, F. J. Activation of Shaker potassium channels. III. An activation gating model for wild-type and V2 mutant channels. *J. Gen. Physiol.* **111**, 313–342 (1998).

18. Armstrong, C. M. Interaction of tetraethylammonium ion derivatives with the potassium channels of giant axons. *J. Gen. Physiol.* **58**, 413–437 (1971).
19. Hoshi, T., Zagotta, W. N. & Aldrich, R. W. Biophysical and molecular mechanisms of Shaker potassium channel inactivation. *Science* **250**, 533–538 (1990).
20. Hoshi, T., Zagotta, W. N. & Aldrich, R. W. Two types of inactivation in Shaker K⁺ channels: effects of alterations in the carboxy-terminal region. *Neuron* **7**, 547–556 (1991).
21. Swartz, K. J. & MacKinnon, R. Hanatoxin modifies the gating of a voltage-dependent K⁺ channel through multiple binding sites. *Neuron* **18**, 665–673 (1997).
22. Swartz, K. J. & MacKinnon, R. Mapping the receptor site for hanatoxin, a gating modifier of voltage-dependent K⁺ channels. *Neuron* **18**, 675–682 (1997).
23. McDonough, S. I., Lampe, R. A., Keith, R. A. & Bean, B. P. Voltage-dependent inhibition of N- and P-type calcium channels by the peptide toxin omega-granulotoxin-SIA. *Mol. Pharmacol.* **52**, 1095–1104 (1997).
24. Li-Smerin, Y. & Swartz, K. J. Gating modifier toxins reveal a conserved structural motif in voltage-gated Ca²⁺ and K⁺ channels. *Proc. Natl Acad. Sci. USA* **95**, 8585–8589 (1998).
25. Middleton, R. E. *et al.* Two tarantula peptides inhibit activation of multiple sodium channels. *Biochemistry* **41**, 14734–14747 (2002).
26. Heginbotham, L., LeMasurier, M., Kolmakova-Partensky, L. & Miller, C. Single streptomyces lividans K⁺ channels. Functional asymmetries and sidedness of proton activation. *J. Gen. Physiol.* **114**, 551–560 (1999).
27. Park, C. S., Hausdorff, S. F. & Miller, C. Design, synthesis, and functional expression of a gene for charybdotoxin, a peptide blocker of K⁺ channels. *Proc. Natl Acad. Sci. USA* **88**, 2046–2050 (1991).
28. Gill, S. C. & von Hippel, P. H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **182**, 319–326 (1989).
29. Diochot, S., Drici, M. D., Moinier, D., Fink, M. & Lazdunski, M. Effects of phrixotoxins on the Kv4 family of potassium channels and implications for the role of Ito1 in cardiac electrogenesis. *Br. J. Pharmacol.* **126**, 251–263 (1999).
30. Escoubas, P., Diochot, S., Celerier, M. L., Nakajima, T. & Lazdunski, M. Novel tarantula toxins for subtypes of voltage-dependent potassium channels in the Kv2 and Kv4 subfamilies. *Mol. Pharmacol.* **62**, 48–57 (2002).

Acknowledgements We thank M. Trester-Zedlitz for assistance with mass spectrometry and W. Chin for assistance in manuscript preparation. Protein sequence was done at the Rockefeller University Protein/DNA Technology Center (supported by NIH shared instrumentation grants, the US Army and Navy). This research was supported by an NIH grant. V.R. is supported by a NSF graduate student research fellowship and R.M. is an Investigator in the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.M. (e-mail: mackinn@rockvax.rockefeller.edu).

Insight into a natural Diels–Alder reaction from the structure of macrophomate synthase

Toyoyuki Ose*, Kenji Watanabe†, Takashi Mie†, Mamoru Honma†, Hiromi Watanabe*, Min Yao*, Hideaki Oikawa† & Isao Tanaka*

* Division of Biological Science, Graduate School of Science, Hokkaido University, Sapporo 060-0810, Japan

† Division of Applied Bioscience, Graduate School of Agriculture, Hokkaido University, Sapporo 060-8589, Japan

The Diels–Alder reaction, which forms a six-membered ring from an alkene (dienophile) and a 1,3-diene, is synthetically very useful for construction of cyclic products with high regio- and stereoselectivity under mild conditions¹. It has been applied to the synthesis of complex pharmaceutical and biologically active compounds². Although evidence^{3–7} on natural Diels–Alderase has been accumulated in the biosynthesis of secondary metabolites⁸, there has been no report on the structural details of the natural Diels–Alderase. The function and catalytic mechanism of the natural Diels–Alderase are of great interest owing to the diversity of molecular skeletons in natural Diels–Alder adducts⁸. Here we present the 1.70 Å resolution crystal structure of the natural Diels–Alderase, fungal macrophomate synthase

(MPS)³, in complex with pyruvate. The active site of the enzyme is large and hydrophobic, contributing amino acid residues that can hydrogen-bond to the substrate 2-pyrone. These data provide information on the catalytic mechanism of MPS, and suggest that the reaction proceeds via a large-scale structural reorganization of the product.

The phytopathogenic fungus *Macrophoma commelinae* has the ability to transform 2-pyrone derivatives into the corresponding benzoate analogues⁹ as macrophomate **1** (Fig. 1). This complex transformation is catalysed by only one enzyme, MPS, with oxalacetate as a substrate for the C₃-unit precursor. MPS is a Mg²⁺-dependent enzyme with 339 amino acid residues (relative molecular mass $M_r = 36,244$)^{10,11}. The catalytic mechanism of the whole pathway has been investigated extensively: the transformation proceeds through three separate steps including decarboxylation of oxalacetate, carbon–carbon bond formations (Diels–Alder reaction), and decarboxylation with concomitant dehydration³. In the absence of 2-pyrone, MPS simply acts as a decarboxylase with high catalytic efficiency (Fig. 1).

The crystal structure of the MPS complexed with pyruvate and Mg²⁺ was determined up to a resolution of 1.70 Å by the multi-

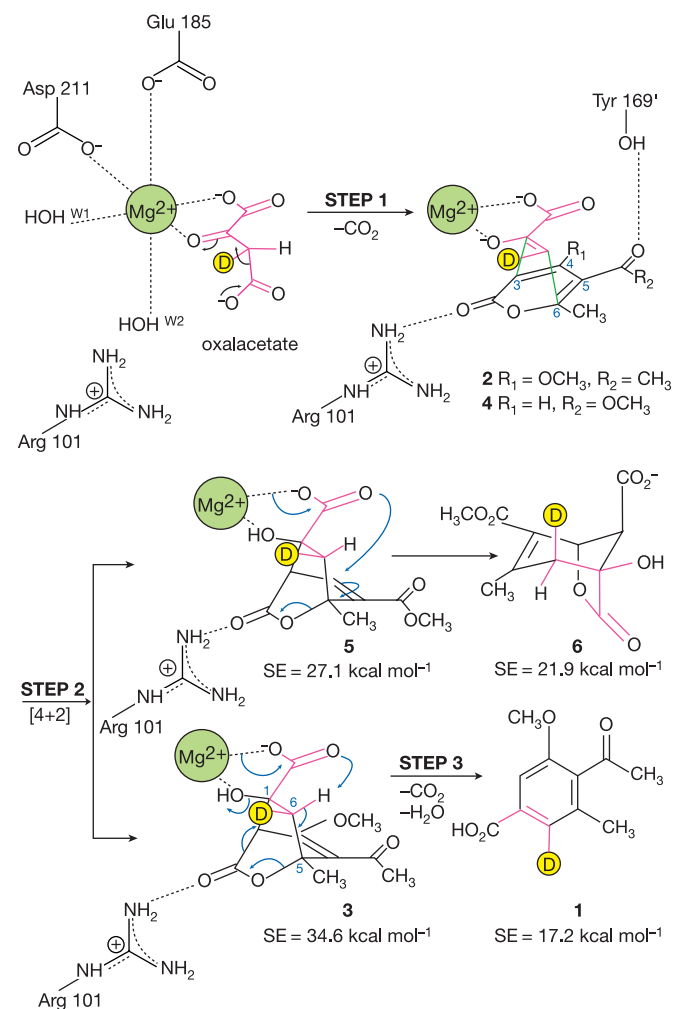


Figure 1 Details of individual reaction steps with macrophomate synthase. Step 1 is decarboxylation of oxalacetate. Step 2 are Diels–Alder reactions of the enolate and 2-pyrones **2** and **4** to form higher energy adducts **3** and **5**, respectively. Step 3 is degradation of **3** in which abstraction of hydrogen triggers C–O bond cleavage followed by decarboxylation and elimination of hydroxy group. The steric energies (SE) of each compound were determined by molecular mechanics calculations using the MM2 force field.

wavelength anomalous diffraction method. The refined structural model of the MPS consists of 299 residues out of a total of 339 residues of the polypeptide chain. The C-terminal 40 residues (residues 300–339) were not constructed, because the electron density is poor in this region. The deletion mutant of these 40 residues showed the same MPS-specific activity as the native enzyme (data not shown), suggesting that this region is not directly involved in the catalytic activity. The molecule is hexameric with a point group symmetry of 32. The protomer core region consists of an eight-stranded β -barrel surrounded by 8 + 3 α -helices with a $(\beta/\alpha)_8$ barrel fold; the C-terminal $\alpha 8$ -helix (residues 275–298) of each protomer protrudes from the core and joins to the $(\beta/\alpha)_8$ barrel of the 2-fold-related protomer (Fig. 2a). With these swapped helices, two protomers are closely associated to form an extensively hydrophobic dimer interface (Fig. 2b). Three of these dimers associate further to form the hexameric molecule with the active site being at the interface with a three-fold-related protomer (Fig. 2c, d). The tertiary as well as the quaternary structure of the MPS described above is similar to that of the 2-dehydro-3-deoxygalactarate (DDG) aldolase¹². DDG aldolase catalyses the reversible aldol reaction of DDG to pyruvate and tartronic semialdehyde, whereas MPS catalyses multi-step conversion reactions including two carbon–carbon bond formations. Thus, the catalytic activities

of these two enzymes are very different. More detailed comparisons of the two enzymes will be described elsewhere.

The active site of the MPS is at the C-terminal end of the β -barrel, which is covered by the long loop from the three-fold-related chain. At one side of this catalytic cavity, Mg^{2+} is located in an octahedral coordination (Fig. 3a). Two of the ligands of Mg^{2+} are side-chain carboxyl oxygens from residues Glu 185 and Asp 211. Another two coordination sites are filled with two water molecules, W1 and W2, which are in turn hydrogen-bonded to the amino-acid residues from the three-fold-related protomer. The last two coordination sites are occupied by the C2-carbonyl and C1-carboxyl oxygen atoms of pyruvate. The high-resolution electron-density map of the present analysis clearly defines the precise orientation of the pyruvate (Fig. 3b), of which two carboxyl oxygen atoms are also hydrogen-bonded to main-chain amide protons of Gly 210 and Asp 211. This complex is further stabilized by interaction between the carbonyl oxygen of pyruvate and side-chains of Arg 101. With these bonds, the pyruvate is tightly placed at this position. Underneath the pyruvate binding site, there is a space where the second substrate 2-pyrone can bind. This space is hydrophobic and large enough to accommodate 2-pyrone and to allow necessary conformational change owing to the Diels–Alder reaction.

For the first decarboxylation reaction, oxalacetate needs to be incorporated into the active site of MPS to form a chelation complex similar to that of pyruvate (Fig. 3a). Lewis acidity of the magnesium promotes decarboxylation from oxalacetate to form the enolate anion, which is stabilized by an electron sink provided from the divalent cation. This model is consistent with the stereochemical course of the decarboxylation predicted by the previous experiment with (3S)-[3-²H]-oxalacetate³.

In the second step of the reaction, the cycloaddition of the enolate and the 2-pyrone 2 takes place (Fig. 1). As will be described here, we have already obtained unambiguous evidence on the structure and the absolute configuration of the initial Diels–Alder reaction product 3 (ref. 3). This provides definitive information on the second substrate (2-pyrone) in the catalytic site in terms of the approaching direction and orientation. On the basis of the structure of 3 and the present crystal structure, we built a model for the very early transition state of the Diels–Alder reaction as shown in Fig. 3c. In this binding model, the 2-pyrone molecule is likely to be fixed in place through two hydrogen bonds between the carbonyl oxygen of 2-pyrone and Arg 101, and the C5-acyl oxygen and Tyr 169. Tyr 169 is in turn placed in the proper orientation through stacking with Phe 149. The flexible loop (residues 139–170) with hydrophobic side-chains (Phe 149, Pro 151 and Trp 152) from the three-fold-related protomer shields this transition state from the solvent. To determine the importance of the hydrogen bonds in the present model, we prepared two mutants, R101S and Y169F. Both mutations greatly disturbed MPS activity while retaining the decarboxylase activity (no MPS activity remains for R101S and the V_{max} value of Y169F is below 10% (from 0.87 to 0.07 $\mu\text{mol min}^{-1} \text{mg}^{-1}$) with almost the same K_m value), confirming the importance of these hydrogen bonds in the carbon–carbon bond-forming reaction and providing strong support for the proposed model in Fig. 3c. (V_{max} is the maximum velocity of the MPS-catalysed reaction, and K_m is the Michaelis constant for decarboxylation of oxalacetate.)

Our binding model (Fig. 3c) shows that there is enough space around the C4, C5 and C6 positions of 2-pyrone, but the free space around the C3 is limited owing to the presence of Mg^{2+} and its ligands. This environment surrounding the substrate ties in well with the substrate specificity of MPS: various bulky groups can be substituted on the C4, C5 and C6 positions, but no substitution is allowed on the C3 position¹³. The experimental result that 2-pyrones lacking a C5-acyl group are not converted into normal aromatic products¹³ gives further support for this binding structure.

The final chemical step is the elimination of water and carbon

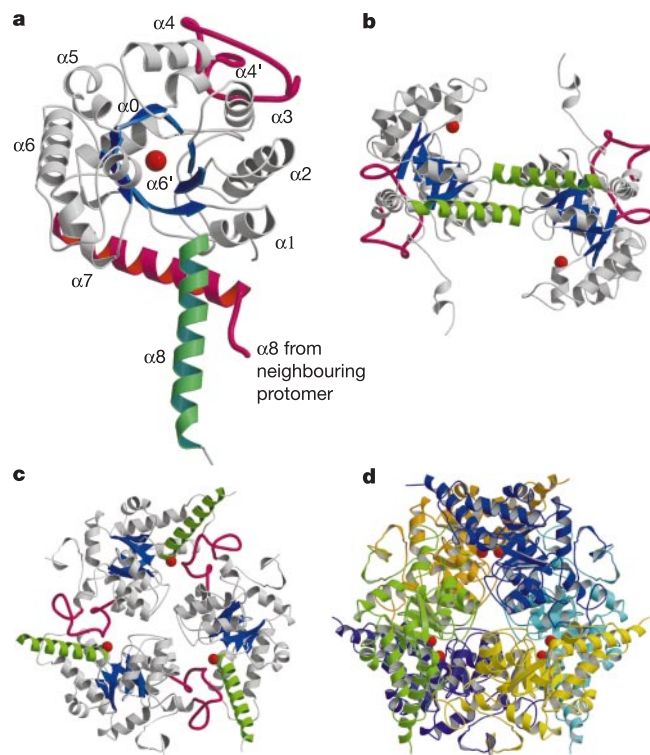


Figure 2 Overall structure of MPS. **a**, Protomer structure of MPS showing α -helix swapped $(\beta/\alpha)_8$ barrel fold. The core β -barrel (blue) is surrounded by 11 α -helices. The long $\alpha 8$ -helix (coloured in magenta) belongs to the neighbouring protomer related by the two-fold axis and joins to the β -barrel core to form a complete $(\beta/\alpha)_8$ barrel. The divalent magnesium ion (red ball) is located at the C-terminal region of the β -barrel. A long loop (coloured in magenta) opposite to the swapped helix interacts with another adjacent protomer related by the three-fold axis (as shown in **c**). **b**, Dimer structure formed by helix swapping. Two protomers related by the two-fold axis are tightly associated by the hydrophobic interactions of the swapped helices (drawn in green). **c**, Three protomers related by the three-fold axis. Each flexible loop coloured in magenta (residues 139–171) joins to the active site of the next protomer. The flexibility of this region may help substrates migrate into catalytic pockets. **d**, The functional unit of MPS with point group symmetry 32.

dioxide (Fig. 1). Previously, we found that the 2-pyrone **4** lacking the C4-methoxy group rapidly reacted to give aberrant adduct **6** (refs 3, 14). Thus, there are two routes for transforming higher energy $[4 + 2]$ adducts **3** and **5** to either the benzoate analogue **1** or the rearranged product **6**. In the reaction with the normal 2-pyrone **2**, substitution of the electron-donating methoxy group reduces the reactivity of the α , β -unsaturated ester, with the result that the weak nucleophile—the C1-carboxylate—cannot attack this moiety. This alters the reaction path to form the aromatic compound **1**.

The representative route of the degradation step from the adduct **3** to **1** is shown in Fig. 1. Because dehydration formally proceeds through *anti*-elimination, on the basis of our observations³, this step involves deprotonation of the pro-*R* proton at C6 and decarboxylation cleaving the C5–O bond antiperiplanar to the pro-*R* proton concomitant with elimination of the hydroxy group. Kinetic analysis of MPS for the first decarboxylation (k_{cat} is 16.3 s^{-1}), the formation of the aberrant adduct **6** (k_{cat} is 5.9 s^{-1}), and the overall reaction (k_{cat} is 0.6 s^{-1}) reveals that the last degradation step is a rate-determining step³. The catalytic inefficiency of the final degradation step may be attributed to intramolecular deprotonation of pro-*R* hydrogen with the C1-carboxyl group in intermediate **3**.

At present, three examples of natural Diels–Alderase are known (Fig. 4). Solanapyrone synthase^{4–6}, lovastatin nonaketide synthase⁷ and MPS catalyse not only the Diels–Alder reaction but also oxidation, polyketide chain formation and decarboxylation, respectively. These enzymes convert the corresponding substrates into reactive Diels–Alder substrates, which are not released from the active sites and readily undergo cycloaddition in the active sites, forcing them into reactive conformation. Thus, among the over a hundred natural plausible Diels–Alderase⁸, the three examples shown above can at least be classified as producers of reactive substrate with an entropy trap for $[4 + 2]$ cycloaddition. Compared

with artificial biomolecular catalysts, this type of natural Diels–Alderase has an obvious advantage for its ability to use a highly reactive substrate that is not stable in reaction medium.

Macrophomate ester was effectively synthesized from ethyl proliolate and the 2-pyrone¹⁵. This type of cycloaddition between an electron-rich dienophile and an electron-deficient diene, such as a 2-pyrone, is categorized as an inverse-electron-demand Diels–Alder reaction, which normally provides an aromatic compound¹⁶. In this reaction, the electron-withdrawing groups at C3 and C5 of the 2-pyrone increase the reactivity of diene and, moreover, the coordination of Lewis acid^{16,17} or the hydrogen bonding¹⁸ of solvent and catalytic residues^{19–21} in the active site with the carbonyl groups effectively decreases the lowest unoccupied molecular orbital (LUMO) energy of the diene, resulting in an increase of the reactivity. In the MPS reaction, two hydrogen bonds and the appropriate orientation of the enolate and the 2-pyrone to maximize the overlapping of π -orbitals in the dienophile and the diene effectively promote the Diels–Alder reaction to produce a higher-energy adduct **3**. Although all the evidence suggests that the MPS reaction is a Diels–Alder reaction, proof that this is a concerted reaction is still required to exclude the stepwise Michael–aldol reaction, described previously as an alternative reaction mechanism of the carbon–carbon bond formations in the MPS reaction³.

The Diels–Alder reaction usually does not require any catalytic process¹⁷, and thus rate acceleration can easily be achieved by entropic factors. In fact, it has been reported that the antibody that is elicited by the transition state analogue is able to catalyse the Diels–Alder reaction^{22–25}. Product inhibition is an inherent problem in the reaction with Diels–Alderase antibody because of the resemblance between the product and the transition state. Hilvert and co-workers developed Diels–Alderase antibody 1E9 (refs 22, 26), in

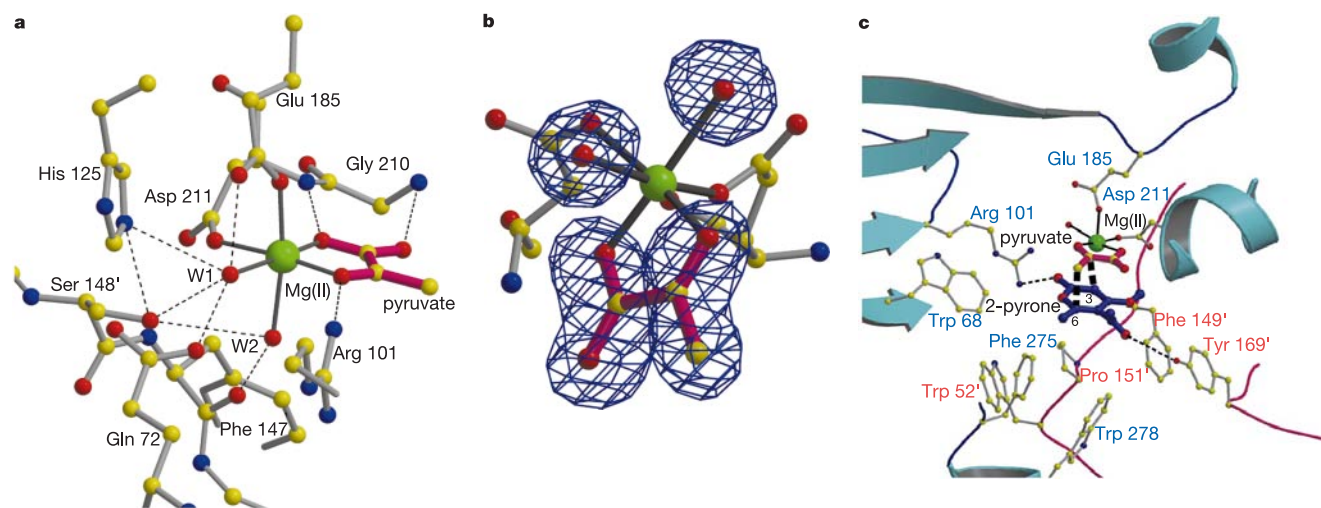


Figure 3 Active site view showing Mg^{2+} coordination, the electron density map, and the proposed model of the early transition state. **a**, Schematic representation depicting the detailed interactions of octahedrally coordinated Mg^{2+} (green ball) and its ligands. The ligands are the carboxyl oxygens of Glu 185 and Asp 211, C2-carbonyl and C1-carboxyl oxygen atoms of pyruvate, and two water molecules. Profound networks of the hydrogen bonds are shown (dotted lines) around these ligands. The bonds and atoms are coloured as follows: pyruvate moiety (magenta), MPS (grey), coordination bonds of Mg^{2+} (black), oxygen (red), nitrogen (blue), carbon (yellow) and magnesium (green). **b**, The final σ_A weighted $F_o - F_c$ omit electron density for pyruvate and two water molecules. The contour level of the electron density map is 3.0σ and the resolution is $1.70 \text{ \AA}$. **c**, The residues in the active site pocket and proposed model for the very early transition state of the Diels–Alder reaction. 2-pyrone **2** (blue thick bond) is placed in parallel with pyruvate enolate (magenta thick bond) which is bound to Mg^{2+} (green ball). C2 carbonyl and C5 acyl-

carbonyl oxygens of 2-pyrone **2** are located respectively within hydrogen-bonding distance to guanidyl nitrogen of Arg 101 and phenol oxygen of Tyr 169' (here the prime sign represents the residue from the neighbouring chain). Side chains of Arg 101 and Tyr 169' are drawn as they are found in the crystal and could slightly adjust their conformations for making hydrogen bonds to 2-pyrone **2**. The hydrogen bonds are drawn as thin dotted lines and the incoming carbon–carbon bonds between C2 (and C3) of pyruvate and C3 (and C6) of 2-pyrone **2** are drawn as thick dotted lines. The wall of the active site pocket at the 2-pyrone side consists of mostly hydrophobic residues with bulky side chains such as Trp 68, Phe 275, Trp 278, Phe 149', Pro 151' and Trp 152'. There are no ionic residues except Arg 101, and Tyr 169' (mentioned above) in the active site close to the 2-pyrone **2**. Atoms are coloured as in **a**. The main-chain trace drawn in red is of the neighbouring protomer related by the three-fold axis.

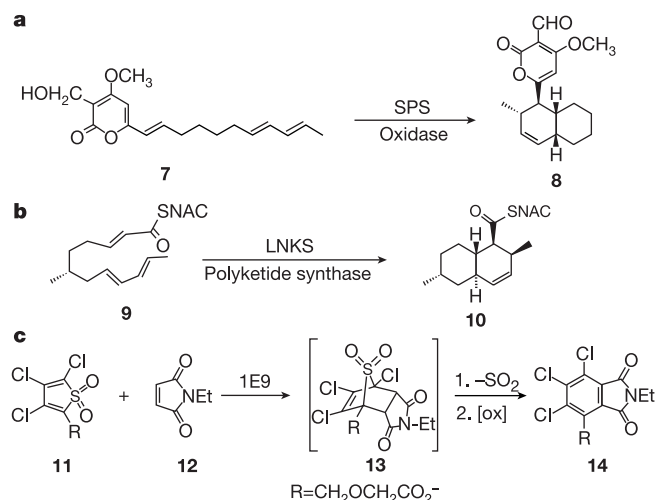


Figure 4 Comparison of Diels-Alderase. **a**, Solanapyrone synthase (SPS) catalyses oxidation of alcohol **7** to the reactive formyl derivative which readily promotes [4 + 2] cycloaddition to give solanapyrone A **8**. **b**, Lovastatin nonaketide synthase (LNKS) catalyses intramolecular [4 + 2] cycloaddition from **9** to **10**. LNKS also catalyses condensation of acetyl CoA and malonyl CoA to form an enzyme bound analogue of **9**. SNAC is N-acetylcysteamine thioester. **c**, Diels-Alderase antibody 1E9 transforms thiophene dioxide **11** and maleimide **12** to intermediate **13** via [4 + 2] cycloaddition, which is then converted into aromatic product **14** with elimination of sulphur dioxide and the subsequent oxidation. Non-catalysed degradation from **13** to **14** allows this catalytic antibody to escape the product inhibition. This leads 1E9 to be the most efficient Diels-Alderase antibody.

which uncatalysed SO₂ elimination from the product is programmed so as to avoid product inhibition (Fig. 4).

The first structure analysis of the natural Diels-Alderase allows a comparative study between natural (MPS) and artificial Diels-Alderase (1E9; ref. 26). Both biocatalysts adopt a number of common strategies for the reaction: possessing an entropy trap to stabilize the transition state, programming the elimination (CO₂, SO₂) from the initial Diels-Alder adduct to avoid product inhibition, and use of the hydrogen bonds (in MPS, between Arg 101, Tyr 169 and carbonyl groups of the 2-pyrone; in 1E9 (ref. 26), between Asn^H35 and the carbonyl group of **12**) for recognition and activation of the substrate. Both Diels-Alderase have an optimized hydrophobic cavity that allows large structural change of the substrate/product during the Diels-Alder reaction. However, in substrate recognition and the rate-determining step there are large differences between 1E9 and MPS. In the case of 1E9, interactions of the substrates with the active site occur mostly through van der Waals contacts²¹ whereas electrostatic interactions and a number of hydrogen bondings are also observed in the case of MPS. These differences may contribute to the difference in affinity of substrates to the corresponding Diels-Alderase (MPS (decarboxylation): $K_m = 1.2 \text{ mM}$ ($K_i = 0.09 \text{ mM}$) for oxalacetate, $K_m = 1.6 \text{ mM}$ ($K_i = 0.16 \text{ mM}$) for **2**; 1E9: $K_m = 29 \text{ mM}$ for **12**, 2.4 mM for **11**). Kinetic analysis of 1E9 shows that the rate-limiting step is cycloaddition, while that in MPS is degradation of the adduct; MPS exhibits higher catalytic efficiency ($k_{cat}/K_m = 500 \text{ M}^{-1} \text{ s}^{-1}$ for oxalacetate, $375 \text{ M}^{-1} \text{ s}^{-1}$ for **2**) than does 1E9 ($k_{cat}/K_m = 7.5 \text{ M}^{-1} \text{ s}^{-1}$ for **12**, $90 \text{ M}^{-1} \text{ s}^{-1}$ for **11**). In addition, MPS has the flexible loop, which may function to incorporate a substrate and to expel a product. (K_i is the inhibition constant for compound **2** in the decarboxylation reaction catalysed by MPS.)

Thus the crystal structure of MPS provides information regarding the active site interaction among the substrates, metal and amino acid residues of MPS, which are essential for the catalysis.

These data provide the catalytic mechanism of the complex multiple reaction and the Diels-Alder reaction escaping product inhibition. These findings could help in the design of a Diels-Alderase antibody and also in the modification of the existing enzyme to Diels-Alderase. □

Methods

Data collection and processing

The overexpression, purification, and crystallization of native and selenomethionine-substituted MPS will be described elsewhere. The crystals of native and selenomethionine-substituted MPS were obtained by the hanging-drop vapour diffusion method at room temperature. While the crystal of native MPS belongs to space group $P6_322$ with unit cell dimensions of $a = b = 106.7 \text{ Å}$, and $c = 120.8 \text{ Å}$, and diffracts to 1.7 Å resolution, the crystal of selenomethionine-substituted MPS is formed in space group $R32$ with unit cell dimensions of $a = b = 122.6 \text{ Å}$, and $c = 95.8 \text{ Å}$, and diffracts to 2.3 Å resolution. Three-wavelength multiple anomalous diffraction and native data were collected under 100 K on the beamline BL18B of Photon Factory (Tsukuba, Japan). All of the diffraction data sets were integrated and merged using the MOSFLM-SCALA²⁷.

Structure determination and refinement

The multiple anomalous diffraction data were further scaled using the SCALEIT²⁷. An initial model was built from a density map which was calculated with all 11 selenium atom sites and modified with solvent flattening by SHARP/SOLOMON^{27,28} at 2.3 Å resolution. After the model was transferred to the native crystal by molecular replacement using AMoRe²⁷, the molecular dynamic refinement was performed using the CNS²⁹ with a $20\text{--}1.7 \text{ Å}$ resolution region of native data. A pyruvate molecule and a magnesium ion were clearly visible in both σ_A -weighted $2F_o - F_c$ and $F_o - F_c$ density maps, and were included in the latter part of the refinement. The final structure of the MPS-pyruvate complex, which comprises 299 residues, one magnesium ion, one pyruvate molecule and 360 water molecules, was refined to an R_{free} factor of 20.3% and an R factor of 17.6%.

Construction and analysis of MPS mutants

We used the expression vector of MPS as a template to construct a plasmid for the expression of mutant proteins, and introduced site-directed mutation using the TAKARA Mutan-Super Express K_m kit. The result of mutations was confirmed by DNA-sequencing using the Applied Biosystems dRhodamine kit. We transformed expression plasmids carrying each mutation into BL21(DE3) and purified the mutants as described previously¹¹. The assays of oxalacetate decarboxylase activity were carried out using lactate dehydrogenase³ and assays of macrophomate synthase activity were carried out using reversed-phase high-performance liquid chromatography (HPLC)¹³.

Received 2 October 2002; accepted 27 January 2003; doi:10.1038/nature01454.

- Carruthers, W. *Cycloaddition Reactions in Organic Synthesis* (Pergamon, Oxford, 1990).
- Desimoni, G., Tacconi, G., Barco, A. & Pollini, G. P. *Natural Products Synthesis through Percyclic Reactions* (American Chemical Society, Washington, DC 1983).
- Watanabe, K., Mie, T., Ichihara, A., Oikawa, H. & Honma, M. Detailed reaction mechanism of macrophomate synthase. Extraordinary enzyme catalyzing five-step transformation from 2-pyrones to benzoates. *J. Biol. Chem.* **275**, 38393–38401 (2000).
- Oikawa, H., Katayama, K., Suzuki, Y. & Ichihara, A. Enzymatic activity catalysing exo-selective Diels-Alder reaction in solanapyrone biosynthesis. *J. Chem. Soc. Chem. Commun.* 1321–1322 (1995).
- Katayama, K., Kobayashi, T., Oikawa, H., Honma, M. & Ichihara, A. Enzymatic activity and partial purification of solanapyrone synthase: First enzyme catalyzing Diels-Alder reaction. *Biochim. Biophys. Acta* **1384**, 387–395 (1998).
- Oikawa, H., Kobayashi, T., Katayama, K., Suzuki, Y. & Ichihara, A. Total synthesis of (–)-solanapyrone a via enzymatic Diels-Alder reaction of prosolanapyrone. *J. Org. Chem.* **63**, 8748–8756 (1998).
- Audclair, K. et al. Lovastatin nonaketide synthase catalyzes an intramolecular Diels-Alder reaction of a substrate analogue. *J. Am. Chem. Soc.* **122**, 11519–11520 (2000).
- Ichihara, A. & Oikawa, H. in *Comprehensive Natural Products Chemistry* (eds Barton, D., Nakanishi, K. & Meth-Cohn, O.) Vol. 1, 367–408 (Elsevier, Amsterdam, 1999).
- Sakurai, I., Miyajima, H., Akiyama, K., Shimizu, S. & Yamamoto, Y. Studies on metabolites of *Macrophoma commelinae*. IV. Substrate specificity in the biotransformation of 2-pyrones to substituted benzoic acids. *Chem. Pharm. Bull.* **36**, 2003–2011 (1988).
- Oikawa, H. et al. Macrophomate synthase: Unusual enzyme catalyzing multiple reactions from pyrones to benzoates. *Tetrahedr. Lett.* **40**, 6983–6986 (1999).
- Watanabe, K. et al. Macrophomate synthase: Characterization, sequence, and expression in *Escherichia coli* of the novel enzyme catalyzing unusual multistep transformation from 2-pyrones to benzoates. *J. Biochem.* **127**, 467–473 (2000).
- Izard, T. & Blackwell, N. C. Crystal structures of the metal-dependent 2-dehydro-3-deoxy-galactarate aldolase suggest a novel reaction mechanism. *EMBO J.* **19**, 3849–3856 (2000).
- Watanabe, K., Mie, T., Ichihara, A., Oikawa, H. & Honma, M. Substrate diversity of macrophomate synthase catalyzing an unusual multistep transformation from 2-pyrones to benzoates. *Biosci. Biotechnol. Biochem.* **64**, 530–538 (2000).
- Watanabe, K., Mie, T., Ichihara, A., Oikawa, H. & Honma, M. Reaction mechanism of the macrophomate synthase: Experimental evidence on intermediacy of a bicyclic compound. *Tetrahedr. Lett.* **41**, 1443–1446 (2000).
- Ichihara, A., Murakami, K. & Sakamura, S. Synthesis of pyrenocines A, B and pyrenocetic acid A. *Tetrahedron* **43**, 5245–5250 (1987).

16. Afarinkia, K., Vinader, V., Nelson, T. D. & Posner, G. H. Diels–Alder cycloadditions of 2-pyrones and 2-pyridones. *Tetrahedron* **48**, 9111–9171 (1992).
17. Sauer, J. & Sustmann, R. Mechanic aspects of Diels–Alder reactions: A critical survey. *Angew. Chem. Int. Edn Engl.* **19**, 779–807 (1980).
18. Blake, J. F. & Jorgensen, W. L. Solvent effects on a Diels–Alder reaction from computer simulations. *J. Am. Chem. Soc.* **113**, 7430–7432 (1991).
19. Romesberg, F. E., Spiller, B., Schultz, P. G. & Stevens, R. C. Immunological origins of binding and catalysis in a Diels–Alderase antibody. *Science* **279**, 1929–1933 (1998).
20. Heine, A. *et al.* An antibody exo Diels–Alderase inhibitor complex at 1.95 angstrom resolution. *Science* **179**, 1934–1940 (1998).
21. Chen, J., Deng, Q., Wang, R., Houk, K. & Hilvert, D. Shape complementarity, binding-site dynamics, and transition state stabilization: A theoretical study of Diels–Alder catalysis by antibody 1E9. *ChemBiochem* **1**, 255–261 (2000).
22. Hilvert, D., Hill, K. W., Nared, K. D. & Auditor, M.-T. Antibody catalysis of a Diels–Alder reaction. *J. Am. Chem. Soc.* **111**, 9261–9262 (1989).
23. Braisted, A. C. & Schultz, P. G. An antibody-catalyzed biomolecular Diels–Alder reaction. *J. Am. Chem. Soc.* **112**, 7430–7431 (1991).
24. Gouverneur, V. E. *et al.* Control of the exo and endo pathways of the Diels–Alder reaction by antibody catalysis. *Science* **262**, 204–208 (1993).
25. Ylikauhaluoma, J. T. *et al.* Anti-metallocene antibodies — a new approach to enantioselective catalysis of the Diels–Alder reaction. *J. Am. Chem. Soc.* **117**, 7041–7047 (1995).
26. Xu, J. *et al.* Evolution of shape complementarity and catalytic efficiency from a primordial antibody template. *Science* **286**, 2345–2348 (1999).
27. Collaborative Computational Project, Number 4. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
28. de La Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276**, 472–494 (1997).
29. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).

Supplementary Information accompanies the paper on *Nature's* website
(<http://www.nature.com/nature>).

Acknowledgements We thank S. Wakatsuki, M. Suzuki and N. Igarashi of the Photon Factory, Japan, for help in data collection at beamline BL18B. This work was supported in part by National Project on Protein Structural and Functional Analyses from the Ministry of Education, Science, Sports and Culture of Japan.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.O. (e-mail: hoik@chem.agr.hokudai.ac.jp) and I.T. (e-mail: tanaka@castor.sci.hokudai.ac.jp). Structure coordinates have been deposited in the Protein Data Bank under accession code 1IZC.

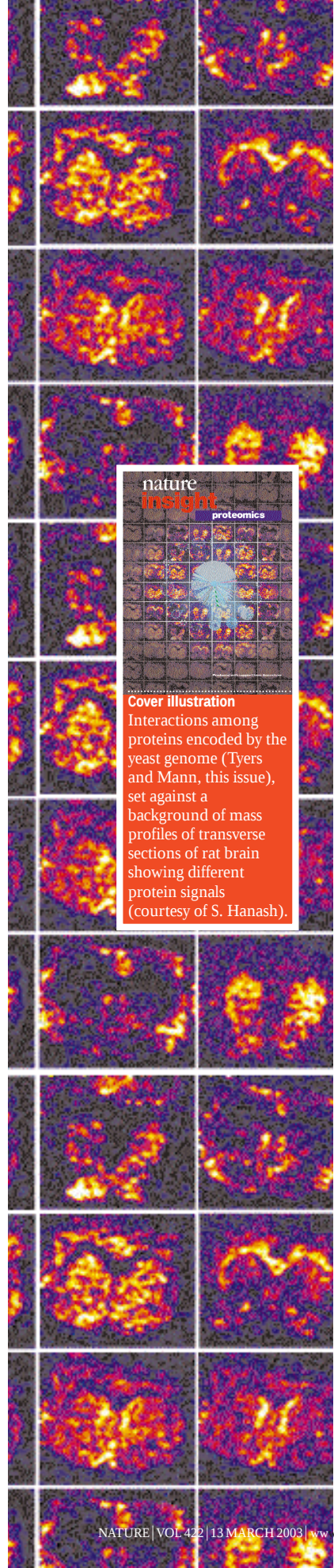
erratum

Attosecond control of electronic processes by intense light fields

A. Baltuška, Th. Udem, M. Uiberacker, M. Hentschel, E. Goulielmakis, Ch. Gohle, R. Holzwarth, V. S. Yakovlev, A. Scrinzi, T. W. Hänsch & F. Krausz

Nature **421**, 611–615 (2003).

In Fig. 2c of this Letter, the units of time on the *x* axis should be seconds, not femtoseconds. □



nature insight

proteomics

**nature
insight
proteomics**

Cover illustration
Interactions among proteins encoded by the yeast genome (Tyers and Mann, this issue), set against a background of mass profiles of transverse sections of rat brain showing different protein signals (courtesy of S. Hanash).

We are only just beginning to appreciate the power and limitations of the genomics revolution, yet hard on its heels proteomics promises an even more radical transformation of biological and medical research. Encoded proteins carry out most biological functions, and to understand how cells work, one must study what proteins are present, how they interact with each other and what they do.

The term proteome defines the entire protein complement in a given cell, tissue or organism. In its wider sense, proteomics research also assesses protein activities, modifications and localization, and interactions of proteins in complexes. It is very much a technology-driven enterprise, and this collection of reviews reflects the progress made and future developments needed to identify proteins and protein complexes in biological samples comprehensively and quantitatively with both high sensitivity and fidelity.

By studying global patterns of protein content and activity and how these change during development or in response to disease, proteomics research is poised to boost our understanding of systems-level cellular behaviour. Clinical research also hopes to benefit from proteomics by both the identification of new drug targets and the development of new diagnostic markers.

Like genomics, the sheer scale of proteomics research makes it a community effort with the Human Proteome Organisation (HUPO) playing an important role in coordinating proteomics projects worldwide. The wealth of information produced poses challenges for data management, and necessitates publicly accessible databases that use agreed standards to describe protein data, allowing data comparison and integration. Furthermore, the expense and scale of proteomics technologies restricts their access, and solutions must be found that allow the widespread use of proteomics tools. In this spirit, in a commentary published in today's issue of *Nature* (422, 115–116; 2003), Ruedi Aebersold proposes a community-wide strategy that could help shift proteomics research towards a 'browsing mode' of searching through existing information.

We are pleased to acknowledge the financial support of Amersham Biosciences in producing this Insight. As always, *Nature* carries sole responsibility for the editorial content and peer review.

Barbara Marte Senior Editor

overview:

- 193 From genomics to proteomics**
M. Tyers & M. Mann

review articles:

- 198 Mass spectrometry-based proteomics**
R. Aebersold & M. Mann

- 208 Protein analysis on a proteomic scale**

*E. Phizicky,
P. I. H. Bastiaens,
H. Zhu, M. Snyder
& S. Fields*

- 216 From words to literature in structural proteomics**

*A. Sali, R. Glaeser,
T. Earnest
& W. Baumeister*

- 226 Disease proteomics**
S. Hanash

- 233 Biomedical informatics for proteomics**

*M. S. Boguski
& M. W. McIntosh*

Editor, *Nature*: Philip Campbell
Insights Editors: Lesley Anson
 Ursula Weiss
Consultant Editor: Bernd Pulverer
Editorial Assistant: Simon Gibson

Art Director: Majo Xeridat
Layouts: Clifford Saunders
Diagrams: Ann Thomson
 Vicky Askew
 Suzanne Coleman

Production Editor: Simon Gribbin
Production: Sue Gray
Marketing: Claire Aspinall
Sponsorship: Mark Greene

From genomics to proteomics

Mike Tyers* & Matthias Mann†

*Samuel Lunenfeld Research Institute, Mount Sinai Hospital, and Department of Medical Genetics and Microbiology, University of Toronto, Toronto, Canada M5G 1X5 (e-mail: tyers@mshri.on.ca)

†Center for Experimental BioInformatics, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark (e-mail: mann@bmb.sdu.dk)

Proteomics is the study of the function of all expressed proteins. Tremendous progress has been made in the past few years in generating large-scale data sets for protein–protein interactions, organelle composition, protein activity patterns and protein profiles in cancer patients. But further technological improvements, organization of international proteomics projects and open access to results are needed for proteomics to fulfil its potential.

The term proteome was first coined to describe the set of proteins encoded by the genome¹. The study of the proteome, called proteomics, now evokes not only all the proteins in any given cell, but also the set of all protein isoforms and modifications, the interactions between them, the structural description of proteins and their higher-order complexes, and for that matter almost everything 'post-genomic'. In this overview we will use proteomics in an overall sense to mean protein biochemistry on an unprecedented, high-throughput scale. The hope, now being realized, is that this high-throughput biochemistry will contribute at a direct level to a full description of cellular function.

Proteomics complements other functional genomics approaches, including microarray-based expression profiles², systematic phenotypic profiles at the cell and organism level^{3,4}, systematic genetics^{5,6} and small-molecule-based arrays⁷ (Fig. 1). Integration of these data sets through bioinformatics will yield a comprehensive database of gene function that will serve as a powerful reference of protein properties and functions, and a useful tool for the individual researcher to both build and test hypotheses. Moreover, large-scale data sets will be crucial for the emerging field of systems biology⁸.

Challenges and approaches in proteomics

Proteomics would not be possible without the previous achievements of genomics, which provided the 'blueprint' of possible gene products that are the focal point of proteomics studies. Although almost trite, the tasks of proteomics can usefully be contrasted with the huge but straightforward challenges initially facing the genome projects. Unlike the scalable exercise of DNA sequencing, with its attendant enabling technologies such as the polymerase chain reaction and automated sequencing, proteomics must deal with unavoidable problems of limited and variable sample material, sample degradation, vast dynamic range (more than 10⁶-fold for protein abundance alone), a plethora of post-translational modifications, almost boundless tissue, developmental and temporal specificity, and disease and drug perturbations. While proteomics is by definition expected to yield direct biological insights, all of these difficulties render any comprehensive proteomics project an inherently intimidating and often humbling exercise.

In this *Nature* Insight, five central pillars of proteomics research are discussed with an emphasis on technological developments and applications. These areas are mass spectrometry-based proteomics, proteome-wide biochemi-

cal assays, systematic structural biology and imaging techniques, proteome informatics, and clinical applications of proteomics. As is apparent from the reviews, the divisions between these areas are somewhat arbitrary, not least because technological breakthroughs often find immediate application on several fronts. More important, biologically useful insights into protein function often emerge from the combination of different proteomic approaches.

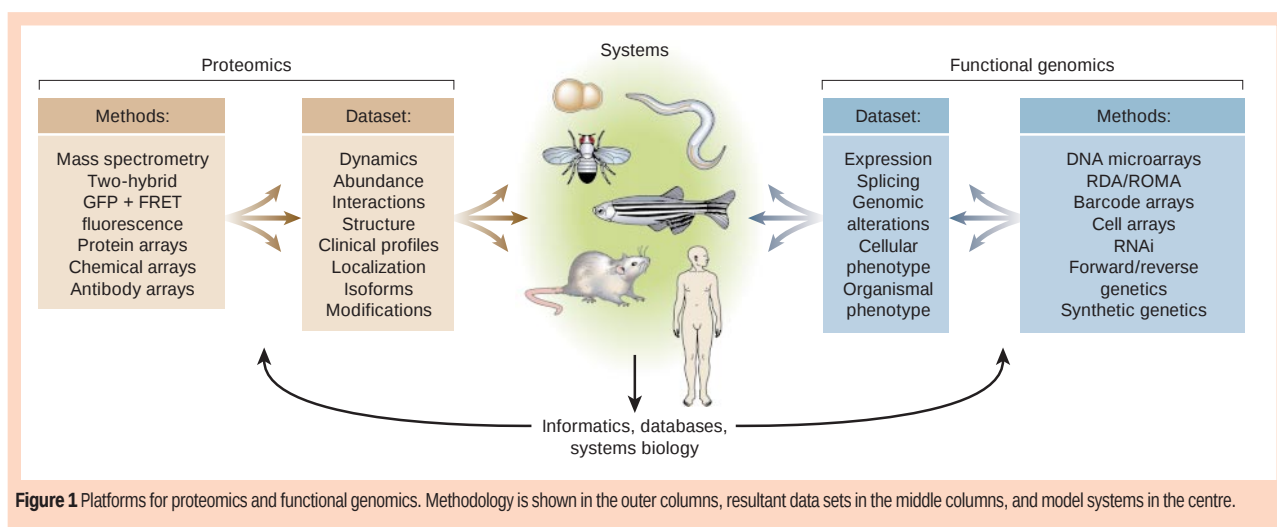
Mass spectrometry-based proteomics

The ability of mass spectrometry to identify ever smaller amounts of protein from increasingly complex mixtures is a primary driving force in proteomics, as described in the review on page 198 by Aebersold and Mann. Initial proteomics efforts relied on protein separation by two-dimensional gel electrophoresis, with subsequent mass spectrometric identification of protein spots. An inherent limitation of this approach is the depth of coverage, which is necessarily constrained to the most abundant proteins in the sample. The rapid developments in mass spectrometry have shifted the balance to direct mass spectrometric analysis, and further developments will increase sensitivity, robustness and data handling.

The past year has seen partial analysis of the yeast interactome, the malaria proteome, bacterial proteomes and various organellar proteomes (see review by Aebersold and Mann, page 198). These vast data sets represent but the tip of the iceberg for biological discovery and drug development. An enormous challenge resides in the obvious fact that the proteome is a dynamic, not a static, entity. Initial efforts to gauge proteome-wide regulatory events in single experiments have been directed at the yeast phosphoproteome⁹ and the ubiquitin-mediated 'degradome' (S. P. Gygi, personal communication). Much higher throughput and sensitivity will be needed to enable true proteome dynamics and moment-by-moment snapshots of cellular responses. Nascent methods for gel-free analysis of complex mixtures hold great promise in this regard¹⁰. Further needs will include more complete sequence coverage of each individual protein, robust and varied methods for sample preparation, and sophisticated algorithms for automated protein identification and detection of post-translational modifications. The ambitious goals of systems biology, which aims to comprehensively model cellular behaviour at the whole-system level^{8,11}, will also require reliable quantitative methods.

Array-based proteomics

A number of established and emergent proteome-wide platforms complement mass spectrometric methods, as



reviewed on page 208 of this issue by Stan Fields and co-workers. The forerunner amongst these efforts is the systematic two-hybrid screen developed by Fields¹². Unlike direct biochemical methods that are constrained by protein abundance, two-hybrid methods can often detect weak interactions between low-abundance proteins, albeit at the expense of false positives.

More recently, various protein-array formats promise to allow rapid interrogation of protein activity on a proteomic scale. These arrays may be based on either recombinant proteins or, conversely, reagents that interact specifically with proteins, including antibodies, peptides and small molecules¹³. Readouts for protein-based arrays can derive from protein interactions, protein modifications or enzymatic activities. A current challenge is to effectively couple high-end mass spectrometry to array formats. Array-based approaches can also use *in vivo* readouts, for example in the systematic analysis of protein localization in the cell through green fluorescent protein (GFP) signals or protein association through fluorescence resonance energy transfer (FRET) between protein fusions to different wavelength variants of GFP. Finally, cell- and tissue-based arrays enable yet another layer of functional interrogation.

One practical bottleneck to these approaches, and indeed to most systematic approaches, has been the limited availability of validated genome-wide complementary DNA for use in the capture of protein complexes with epitope tags. The FlexGene consortium between academic institutions and industry aims to develop complete cDNA collections in recombination-based cloning formats for the biomedical community (see <http://www.hip.harvard.edu>).

Structural proteomics

Beyond a description of protein primary structure, abundance and activities, the ambitious goal of systematically understanding the structural basis for protein interactions and function is reviewed by Baumeister *et al.* on page 216 of this issue. Through literary metaphor, the authors make a compelling argument that a full description of cell behaviour necessitates structural information at the level not only of all single proteins, but of all salient protein complexes and the organization of such complexes at a cellular scale. This all-encompassing structural endeavour spans several orders of magnitude in measurement scale and requires a battery of structural techniques, from X-ray crystallography and nuclear magnetic resonance (NMR) at the protein level, to electron microscopy of mega-complexes and electron tomography for high-resolution visualization of the entire cellular milieu. The recurrent proteomic theme of throughput and sensitivity runs through each of these structural methods, and Baumeister *et al.* suggest novel solutions, even including eliminating the crystals from crystallography! NMR and *in*

silico docking will be necessary to build in dynamics of protein interactions, much of which may be controlled through largely unstructured regions¹⁴.

Informatics

As with any data-rich enterprise, informatics issues loom large on several proteomics fronts. On page 233 of this issue, Boguski and McIntosh highlight the importance of sample documentation, the implementation of rigorous standards and proper annotation of gene function¹⁵. It is crucial that software development is linked at an early stage through agreed documentation, XML-based definitions and controlled vocabularies that allow different tools to exchange primary data sets. Considerable effort has already gone into interaction databases¹⁶ and systems biology software infrastructure¹⁷ that should be built upon by future proteomics initiatives. The development of statistically sound methods for assignment of protein identity from incomplete mass spectral data will be critical for automated deposition into databases, which is currently a painstaking manual and error-prone process. Lessons learned from analysis of DNA microarray data, including clustering, compendium and pattern-matching approaches, should be transportable to proteomic analysis², and it is encouraging that the European Bioinformatics Institute and the Human Proteome Organisation (HUPO) have together started an initiative on the exchange of protein-protein interaction and other proteomic data (see <http://psidev.sourceforge.net/>)

Clinical proteomics

Proteomics is set to have a profound impact on clinical diagnosis and drug discovery, as is fittingly reviewed by Sam Hanash on page 226, the inaugural president of HUPO. Because most drug targets are proteins, it is inescapable that proteomics will enable drug discovery, development and clinical practice. The form(s) in which proteomics will best fulfil this mandate is in a state of flux owing to a multitude of factors, not the least of which are the varied technological platforms in different stages of implementation.

The detection of protein profiles associated with disease states dates back to the very beginning of proteomics, when two-dimensional gel electrophoresis was first applied to clinical material. The advent of mass spectrometers now able to resolve many tens of thousands of protein and peptide species in body fluids is set to revolutionize protein-based diagnostics, as demonstrated in recent retrospective studies of cancer patients¹⁸. The robust and high-throughput nature of mass spectrometric instrumentation is imminently suited to clinical applications. Protein- and antibody-based arrays with validated diagnostic readouts may also become amenable to the clinical

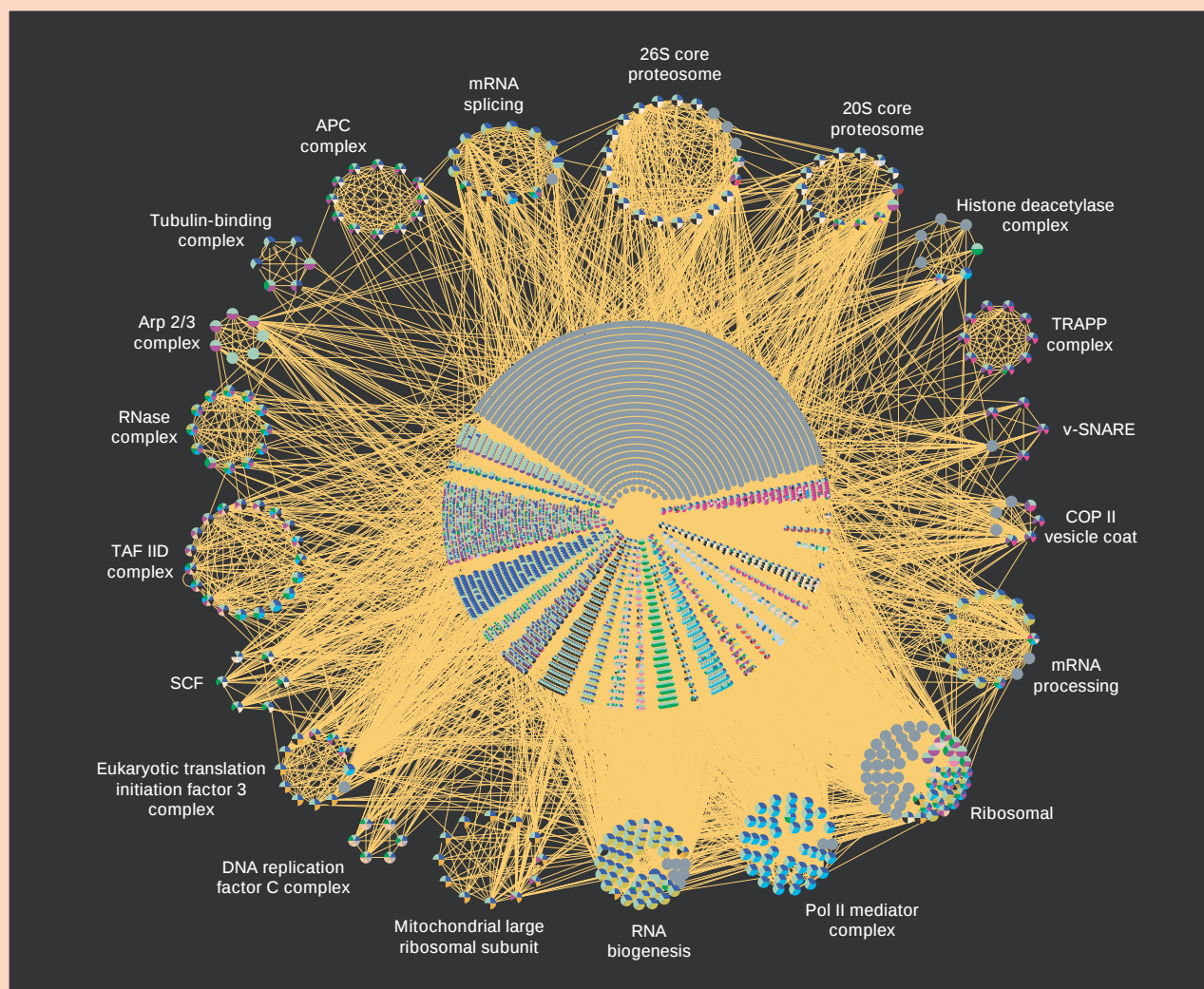


Figure 2 Visualization of combined, large-scale interaction data sets in yeast. A total of 14,000 physical interactions obtained from the GRID database were represented with the Osprey network visualization system (see <http://biodata.mshri.on.ca/grid>). Each edge in the graph represents an interaction between nodes, which are coloured according to Gene Ontology (GO) functional annotation. Highly connected complexes within the data set, shown at the perimeter of the central mass, are built from nodes that share at least three interactions within other complex members. The complete graph contains 4,543 nodes of ~6,000 proteins encoded by the yeast genome, 12,843 interactions and an average connectivity of 2.82 per node. The 20 highly connected complexes contain 340 genes, 1,835 connections and an average connectivity of 5.39.

setting. As with all clinical interfaces, issues of standardized sample preparation, storage and annotation must be addressed.

Proteomics will inevitably accelerate drug discovery, although the pace of progress in this area has been slower than was initially envisaged. Identification of new disease-specific targets, often those present on the cell surface, has been greatly enabled with current technology. An understanding of the biological networks that lie below the cell's exterior will provide a rational basis for preliminary decisions on target suitability.

Orthogonal omics

A caveat of all high-throughput approaches, including proteomics, is that the very scale of experimentation often precludes repetition and rigorous confirmation that is the essence of sound research. However, the intersection between proteomic data sets from different species or between proteomic and other genome-wide data sets often allows robust cross-validation (Fig. 1). This point is aptly illustrated by recent proteomic analysis of the yeast and human nucleolus, in which both directed and undirected efforts uncovered a vast network of protein interactions, many of which impinge on the conserved

process of ribosome biogenesis¹⁹. Independent systematic analysis of yeast-cell size mutants (phenomics) and the gene set regulated by one of these size-control genes (transcriptomics) revealed an unanticipated regulatory relationship between ribosome biogenesis and commitment to cell division²⁰.

Similarly, the integration of interactome, phenome and transcriptome data sets has been used to deduce a new regulatory network in the nematode germline²¹. The combined use of physical, phenotypic and expression data sets can generate non-obvious hypotheses that would otherwise not arise from any individual approach. Even with limited data sets, educated guesses can be made based on simple parameters. For example, an algorithm called ScanSite was used to identify tuberous sclerosis complex-1 as a physiologically relevant substrate of protein kinase B (PKB), based solely on the apparent mass by electrophoresis of the phosphorylated species and an abundance of PKB consensus site sequences²². Finally, new information can often be gained by re-investigating known complexes with new methods. For example, three new components of the heavily studied anaphase-promoting complex have recently been found by multidimensional mass spectrometry²³.

With the numerous initiatives to systematically correlate phenotype with loss of gene function in many model organisms including yeast, nematode, fruitfly, zebrafish, mouse and human, the insights gained from the combined use of large-scale cell biological, transcriptional and proteomic data sets should become synergistic as coverage increases. Most recently, the rapid acquisition of phenotypic data by RNA interference methods, with which it is now possible to systematically interrogate the human genome in tissue-culture cells⁶, will greatly accelerate functional discovery when coupled to proteomic data sets.

Future developments and challenges

As the highly successful effort to sequence the human genome has illustrated, faster and cheaper is the inevitable mantra of any large-scale enterprise. This rhetoric applies doubly so to proteomics, although there is far more to proteomics than just throughput. In its absolute sense, the proteome will be as unreachable as the horizon; rather proteomics will coalesce with other technologies in as yet unimagined ways to converge on an accurate description of cellular properties.

By all criteria, current instrumentation is far from optimal, in part because manufacturers have not yet had the necessary lead time to build machines and associated hardware that are perfectly tailored to protein analysis. Mass spectrometry-based proteomics is nowhere near the physical limit of the few ions needed to register a peak and so a huge increase in performance can be expected in the coming years. As refinements are made in next-generation proteomic instruments, it will be possible to monitor many relevant post-translational modifications and protein interactions in ever more complex mixtures²⁴. As one anticipated example of innovation, throughput and coverage could be greatly enabled by storing mass spectrometric signatures of every protein for real-time data-dependent analysis of highly complex mixtures.

At the level of the individual laboratory, there is undoubtedly a huge market for sensitive and affordable bench-top mass spectrometers for routine applications as analytical devices in all aspects of biological research. Developments in robotic sample preparation, alternative readouts for protein interactions, and microfluidics to minimize sample losses will all factor into achieving the goal of delivering high-powered proteomics to the masses. Equally important, availability of reasonably complete sets of expression and antibody reagents for all proteins would improve the speed and scope of both small- and large-scale proteomics.

With regard to the proteomes of even simple model organisms, all indications are that extant interaction maps are far from saturated. As the density of known interactions increases, testable hypotheses should emerge from the data set at an increasing rate, especially in combination with other genome-wide data sets, including predictions from structural data. Once sufficient dynamics data become available to build first-draft models of cellular behaviour, model refinement will require reiteration of proteomic analyses in numerous mutant and drug-treated conditions. If modelling of simple Boolean networks is a guide, the systems-level behaviour of bona fide protein interaction networks is sure to yield some surprises²⁵.

All this information must obviously be presented in a form that can be processed by the human user. To this end, a great deal more effort must be placed on development of visualization tools, including automated integration with other genome-wide data sets (Fig. 2). There is much room here for novel approaches, many of which are likely to come from other fields that are also suffering from information overload. Examples include sophisticated tools for clustering DNA microarray data and multivariate graphical representations that use coloured readouts to highlight overall trends²⁶, as well as the sophisticated, three dimensional interfaces used in modern computer games.

On the clinical front, comprehensive proteomic analysis of small amounts of diseased tissue will facilitate diagnosis and therapeutic

monitoring, particularly as patterns of disease prediction are recognized empirically from large clinical data sets. Application of phosphoproteomic methods to clinical samples promises what may be the most informative and discriminating readout of cellular status, which can then be used to advantage in diagnosis, drug discovery and elucidation of mechanisms of drug action. The proteomics of host–pathogen interactions should also be an area rich in new drug targets. Regardless of the exact format, robust mass spectrometry and protein-array platforms must be moved into clinical medicine to replace the more expensive and less reliable biochemical assays that are the basis of traditional clinical chemistry. Finally, the nascent area of chemoproteomics will not only allow mechanism of action to be discovered for many drugs, but also has the potential to resurrect innumerable failed small molecules that have dire off-target effects of unknown basis. Relatively little investment in well characterized leads hidden in the archives of pharmaceutical companies may leverage huge therapeutic returns.

Open-access proteomics

An all too common refrain of proteomics has been the limited or non-existent access for the individual biomedical researcher. Although virtually all academic centres have a mass spectrometry facility of some sort, lost samples, failed identifications and inadequate throughput are commonplace. In part, these problems represent the teething stages of a complex technology; additional factors are unaffordable equipment costs and a dearth of highly trained personnel to oversee facilities. As a consequence, most breakthroughs and the generation of raw data in proteomics derive from the work of only a handful of technically inclined laboratories. The burden of improving this circumstance falls on instrument manufacturers, proteomics leaders, funding agencies, academic institutions and the individual user alike. National proteome centres have also been proposed as a way to ensure availability of both expertise and equipment²⁷.

The common effort to map and understand the proteome in its various guises can benefit from lessons learned by genome-sequencing consortia. First and foremost, public access to on-line raw data is essential if there is to be sense of participation across the biomedical research community. Agreements similar to the Bermuda guidelines issued at a critical juncture of the genome projects²⁸ that mandate public accessibility and non-patenting of basic proteomic data would facilitate research in both the academic and industrial sectors. Such data should include the primary structure, post-translational modification, localization and protein–protein interaction pattern of all proteins.

It is important that large-scale proteomics efforts are coordinated, both to avoid duplication and to provide strong rationale for funding agencies. These bodies are in principle willing to support proteomics as a way to reap the rewards of the genome projects, but they will have to be presented with clear goals and rationales of how proteomics will build an infrastructure to advance biomedical science. HUPO is one body that is positioned to play an important coordinating role. HUPO has proclaimed five initial goals for world-wide proteomics research: definition of the plasma proteome, proposals for an in-depth proteomics assault on specific cell types, formation of a consortium to generate antibodies to all human proteins, development of new technologies and formation of an informatics infrastructure. To this list we would add cataloguing the primary structure of all proteins, mapping all organelles that can be purified, and generating protein interaction maps of model organisms, for both comparative proteomics and integration with on-going functional genomics projects.

To meet these laudable goals, it seems that a dedicated funding pool must be established for proteomics research, analogous to that created for the human and model-organism genome sequencing projects, or ongoing funding for these projects should be made available to proteomics. Given the cost of proteomic-scale projects, it

benefits academia and industry to collaborate as much as possible on method development, data acquisition and project coordination. Finally, a way must be established to integrate proteome-scale experiments with efforts of the many individual biology laboratories to develop and test biological models, the final key step in the discovery process that may always defy automation. Whatever the future holds, proteomics will yield great returns for all in what promises to be a knowledge watershed in biology and medicine. □

doi:10.1038/nature01510

1. Wilkins, M. R. *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology* **14**, 61–65 (1996).
2. Shoemaker, D. D. & Linsley, P. S. Recent developments in DNA microarrays. *Curr. Opin. Microbiol.* **5**, 334–337 (2002).
3. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
4. Gerlai, R. Phenomics: fiction or the future? *Trends Neurosci.* **25**, 506–509 (2002).
5. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
6. Hannon, G. J. RNA interference. *Nature* **418**, 244–251 (2002).
7. Kuruvilla, F. G., Shamji, A. F., Sternson, S. M., Hergenrother, P. J. & Schreiber, S. L. Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature* **416**, 653–657 (2002).
8. Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295**, 1664–1669 (2002).
9. Ficarro, S. B. *et al.* Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnol.* **20**, 301–305 (2002).
10. Liu, H., Lin, D. & Yates, J. R. III Multidimensional separations for protein/peptide analysis in the post-genomic era. *Biotechniques* **32**, 898–911 (2002).
11. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
12. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246 (1989).
13. MacBeath, G. Protein microarrays and proteomics. *Nature Genet.* **32**(Suppl.), 526–532 (2002).
14. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
15. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
16. Bader, G. D. & Hogue, W. V. C. in *Genomics and Bioinformatics* (ed. Sensen, C. W.) 399–413 (Wiley-VCH, Weinheim, 2001).
17. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
18. Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C. & Liotta, L. A. Clinical proteomics: translating benchside promise into bedside reality. *Nature Rev. Drug Discov.* **1**, 683–695 (2002).
19. Andersen, J. S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12**, 1–11 (2002).
20. Jorgensen, P., Nishikawa, J. L., Breitkreutz, B. J. & Tyers, M. Systematic identification of pathways that couple cell growth and division in yeast. *Science* **297**, 395–400 (2002).
21. Walhout, A. J. *et al.* Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* **12**, 1952–1958 (2002).
22. Manning, B. D., Tee, A. R., Logsdon, M. N., Blenis, J. & Cantley, L. C. Identification of the tuberous sclerosis complex-2 tumor suppressor gene product tuberlin as a target of the phosphoinositide 3-kinase/akt pathway. *Mol. Cell* **10**, 151–162 (2002).
23. Yoon, H. J. *et al.* Proteomics analysis identifies new components of the fission and budding yeast anaphase-promoting complexes. *Curr. Biol.* **12**, 2048–2054 (2002).
24. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nature Biotechnol.* (in the press).
25. Huang, S. & Ingber, D. E. Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell Res.* **261**, 91–103 (2000).
26. Ball, P. Data visualization: picture this. *Nature* **418**, 11–13 (2002).
27. Aebersold, R. & Watts, J. D. The need for national centers for proteomics. *Nature Biotechnol.* **20**, 651 (2002).
28. Marshall, E. Bermuda rules: community spirit, with teeth. *Science* **291**, 1192 (2001).

Acknowledgements We thank B.-J. Breitkreutz for preparing Fig. 2, D. Figeys and members of the Center for Experimental BioInformatics (CEBI) for critical reading of the manuscript. CEBI is supported by a grant from the Danish Natural Research Foundation.

Mass spectrometry-based proteomics

Ruedi Aebersold* & Matthias Mann†

*Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904, USA (e-mail: raebersold@systemsbiology.org)

†Center for Experimental Bioinformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark (e-mail: mann@bmb.sdu.dk)

Recent successes illustrate the role of mass spectrometry-based proteomics as an indispensable tool for molecular and cellular biology and for the emerging field of systems biology. These include the study of protein–protein interactions via affinity-based isolations on a small and proteome-wide scale, the mapping of numerous organelles, the concurrent description of the malaria parasite genome and proteome, and the generation of quantitative protein profiles from diverse species. The ability of mass spectrometry to identify and, increasingly, to precisely quantify thousands of proteins from complex samples can be expected to impact broadly on biology and medicine.

Proteomics in general deals with the large-scale determination of gene and cellular function directly at the protein level. But as the accompanying articles in this issue describe, the field is a collection of various technical disciplines, all of which contribute to proteomics. These include cell imaging by light and electron microscopy, array and chip experiments, and genetic readout experiments, as exemplified by the yeast two-hybrid assay. Another powerful proteomic approach focuses on the *de novo* analysis of proteins or protein populations isolated from cells or tissues. Such studies typically pose challenges owing to the high degree of complexity of cellular proteomes and the low abundance of many of the proteins, which necessitates highly sensitive analytical techniques. Mass spectrometry (MS) has increasingly become the method of choice for analysis of complex protein samples. MS-based proteomics is a discipline made possible by the availability of gene and genome sequence databases and technical and conceptual advances in many areas, most notably the discovery and development of protein ionization methods, as recognized by the 2002 Nobel prize in chemistry.

Here we survey the state of the field, particularly as it has evolved over the three years since the last review in these pages¹. Already, many of the dreams of the discipline have at least been partly realized. MS-based proteomics has established itself as an indispensable technology to interpret the information encoded in genomes. So far, protein analysis (primary sequence, post-translational modifications (PTMs) or protein–protein interactions) by MS has been most successful when applied to small sets of proteins isolated in specific functional contexts. The systematic analysis of the much larger number of proteins expressed in a cell, an explicit goal of proteomics, is now also rapidly advancing, due mainly to the development of new experimental approaches.

Today, proteomics still remains a multifaceted, rapidly developing and open-ended endeavour. Although it has enjoyed tremendous recent success, proteomics still faces significant technical challenges. Each breakthrough that either allows a new type of measurement or improves the quality of data made by traditional types of measurements expands the range of potential applications of MS to molecular and cellular biology. Indeed, this field is already too expansive for a comprehensive, single review; thus we apologize in advance for the many omissions. However, we do

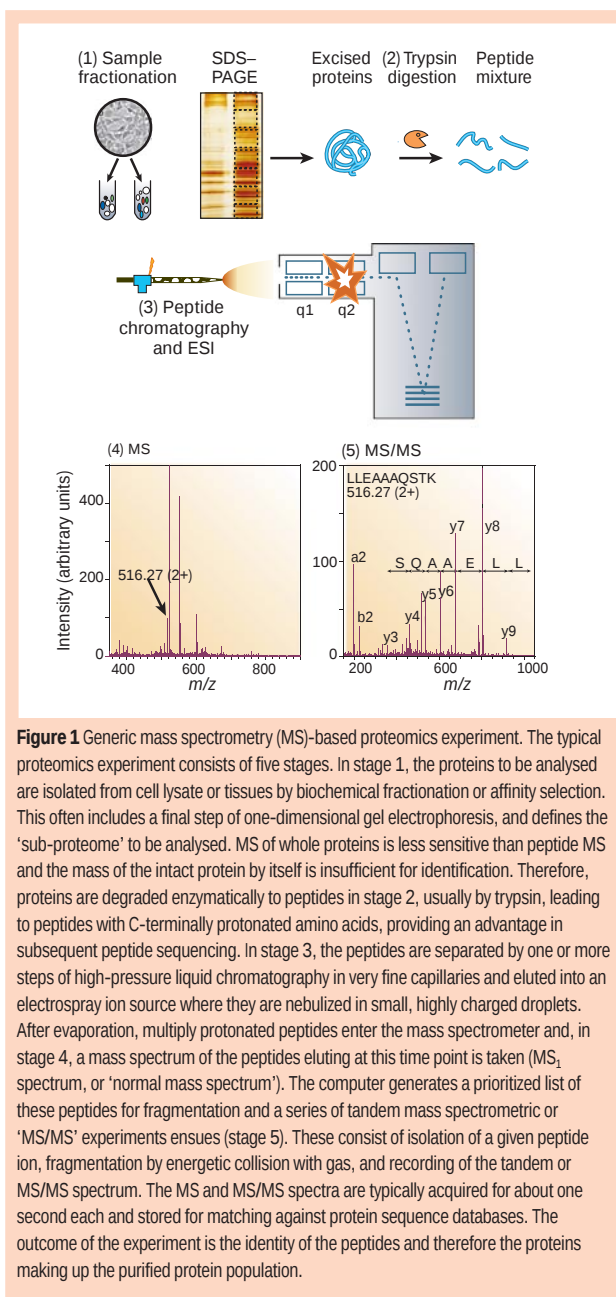
hope that this article captures the excitement of recent achievements in MS-based proteomics, and points the way towards the direction future developments will likely take.

Principles and instrumentation

Mass spectrometric measurements are carried out in the gas phase on ionized analytes. By definition, a mass spectrometer consists of an ion source, a mass analyser that measures the mass-to-charge ratio (m/z) of the ionized analytes, and a detector that registers the number of ions at each m/z value. Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are the two techniques most commonly used to volatilize and ionize the proteins or peptides for mass spectrometric analysis^{2,3}. ESI ionizes the analytes out of a solution and is therefore readily coupled to liquid-based (for example, chromatographic and electrophoretic) separation tools (Fig. 1). MALDI sublimates and ionizes the samples out of a dry, crystalline matrix via laser pulses. MALDI-MS is normally used to analyse relatively simple peptide mixtures, whereas integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex samples.

The mass analyser is, literally and figuratively, central to the technology. In the context of proteomics, its key parameters are sensitivity, resolution, mass accuracy and the ability to generate information-rich ion mass spectra from peptide fragments (tandem mass or MS/MS spectra) (see Fig. 1 and refs 1,4,5). There are four basic types of mass analyser currently used in proteomics research. These are the ion trap, time-of-flight (TOF), quadrupole and Fourier transform ion cyclotron (FT-MS) analysers. They are very different in design and performance, each with its own strength and weakness. These analysers can be stand alone or, in some cases, put together in tandem to take advantage of the strengths of each (Fig. 2).

In ion-trap analysers, the ions are first captured or 'trapped' for a certain time interval and are then subjected to MS or MS/MS analysis. Ion traps are robust, sensitive and relatively inexpensive, and so have produced much of the proteomics data reported in the literature. A disadvantage of ion traps is their relatively low mass accuracy, due in part to the limited number of ions that can be accumulated at their point-like centre before space-charging distorts their distribution and thus the accuracy of the mass measurement. The 'linear' or 'two-dimensional ion trap'^{6,7} is an exciting recent development where ions are stored in a cylindrical volume that is considerably larger than that of



the traditional, three-dimensional ion traps, allowing increased sensitivity, resolution and mass accuracy. The FT-MS instrument is also a trapping mass spectrometer, although it captures the ions under high vacuum in a high magnetic field. Its strengths are high sensitivity, mass accuracy, resolution and dynamic range^{8–11}. But in spite of the enormous potential, the expense, operational complexity and low peptide-fragmentation efficiency of FT-MS instruments has limited their routine use in proteomics research.

MALDI is usually coupled to TOF analysers that measure the mass of intact peptides, whereas ESI has mostly been coupled to ion traps and triple quadrupole instruments and used to generate fragment ion spectra (collision-induced (CID) spectra) of selected precursor ions⁴. More recently, new configurations of ion sources and mass analysers have found wide application for protein analysis. To allow the fragmentation of MALDI-generated precursor ions, MALDI ion sources have recently been coupled to quadrupole ion-trap mass spectrometers¹² and to two types of TOF instruments. In the first, two TOF

sections are separated by a collision cell ('TOF-TOF instrument')¹³, whereas in the second, the hybrid quadrupole TOF instrument, the collision cell is placed between a quadrupole mass filter and a TOF analyser¹⁴. Ions of a particular m/z are selected in a first mass analyser (TOF or quadrupole), fragmented in a collision cell and the fragment ion masses are 'read out' by a TOF analyser. These instruments have high sensitivity, resolution and mass accuracy, and the quadrupole TOF instrument can be used interchangeably with an ESI ionization source. The resulting fragment ion spectra are often more extensive and informative than those generated in trapping instruments. Although TOF, ion-trap and hybrid TOF instruments dominate proteomics today, other configurations including linear ion traps and FT-MS instruments could become widespread in the near future.

As a result of its simplicity, excellent mass accuracy, high resolution and sensitivity, MALDI-TOF is still much used to identify proteins by what is known as peptide mapping, also referred to as peptide-mass mapping or peptide-mass fingerprinting. In this method, proteins are identified by matching a list of experimental peptide masses with the calculated list of all peptide masses of each entry in a database (for example, a comprehensive protein database). Because mass mapping requires an essentially purified target protein, the technique is commonly used in conjunction with prior protein fractionation using either one- or two-dimensional gel electrophoresis (1DE and 2DE, respectively). The addition of sequencing capability to the MALDI method should make protein identifications by MALDI-MS/MS more specific than those obtained by simple peptide-mass mapping (see below). It should also extend the use of MALDI to the analysis of more complex samples, thereby uncoupling MALDI-MS from 2DE. However, if MALDI-MS/MS is to be used with peptide chromatography, the effluent of a liquid chromatography run must be deposited on a sample plate and mixed with the MALDI matrix, a process that has thus far proven difficult to automate. In general, it can be expected that the trend towards the combination of liquid chromatography with ESI- or MALDI-MS/MS (Fig. 1) will continue.

Protein identifications using peptide CID spectra are more clear-cut than those achieved by mass mapping because, in addition to the peptide mass, the peak pattern in the CID spectrum also provides information about peptide sequence. This information is not readily convertible into a full, unambiguous peptide sequence, that is, the 'de novo' sequencing problem via MS is still not generally solved. Instead, the CID spectra are scanned against comprehensive protein sequence databases using one of a number of different algorithms, each with its strengths and weaknesses. The 'peptide sequence tag' approach extracts a short, unambiguous amino acid sequence from the peak pattern that, when combined with the mass information, is a specific probe to determine the origin of the peptide¹⁵. In the cross-correlation method, peptide sequences in the database are used to construct theoretical mass spectra and the overlap or 'cross-correlation' of these predicted spectra with the measured mass spectra determines the best match¹⁶. In the third main approach, 'probability based matching', the calculated fragments from peptide sequences in the database are compared with observed peaks. From this comparison a score is calculated which reflects the statistical significance of the match between the spectrum and the sequences contained in a database¹⁷.

In each of these methods the identified peptides are compiled into a protein 'hit list', which is the output of a typical proteomic experiment. Because protein identifications rely on matches with sequence databases, high-throughput proteomics is currently restricted largely to those species for which comprehensive sequence databases are available.

Protein identification and quantification

No method or instrument exists that is capable of identifying and quantifying the components of a complex protein sample in a simple, single-step operation. Rather, individual components for separating,

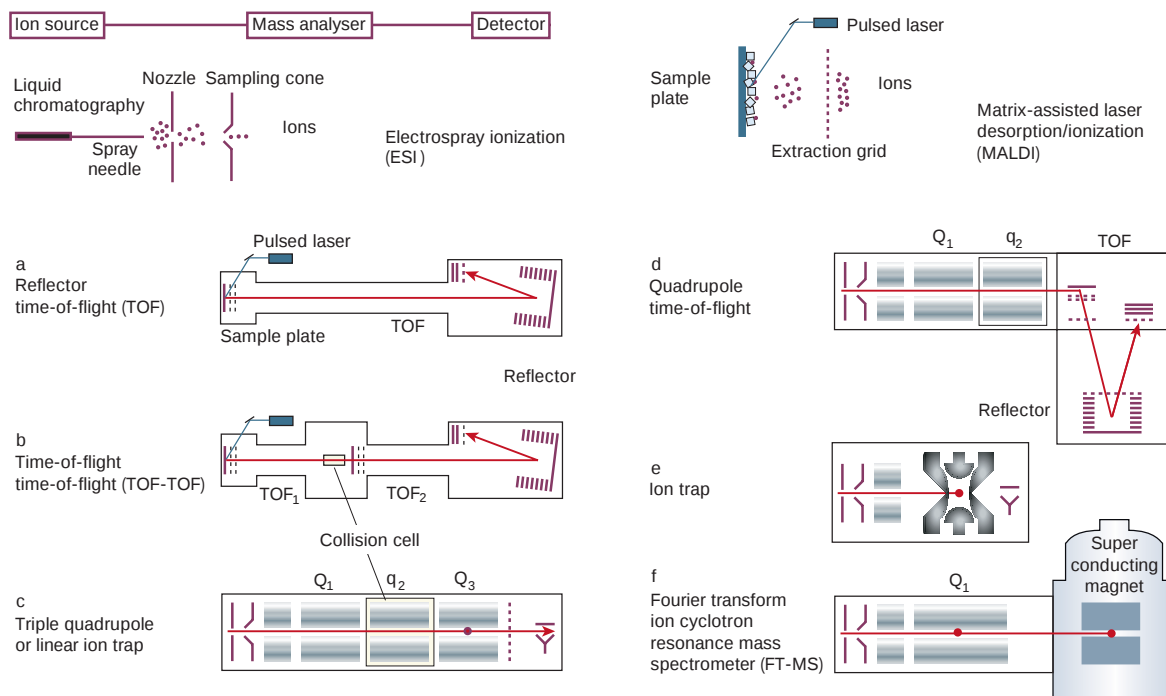


Figure 2 Mass spectrometers used in proteome research. The left and right upper panels depict the ionization and sample introduction process in electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). The different instrumental configurations (a–f) are shown with their typical ion source. **a**, In reflector time-of-flight (TOF) instruments, the ions are accelerated to high kinetic energy and are separated along a flight tube as a result of their different velocities. The ions are turned around in a reflector, which compensates for slight differences in kinetic energy, and then impinge on a detector that amplifies and counts arriving ions. **b**, The TOF-TOF instrument incorporates a collision cell between two TOF sections. Ions of one mass-to-charge (m/z) ratio are selected in the first TOF section, fragmented in the collision cell, and the masses of the fragments are separated in the second TOF section. **c**, Quadrupole mass spectrometers select by time-varying electric fields between four rods, which permit a

stable trajectory only for ions of a particular desired m/z . Again, ions of a particular m/z are selected in a first section (Q_1), fragmented in a collision cell (Q_2), and the fragments separated in Q_3 . In the linear ion trap, ions are captured in a quadrupole section, depicted by the red dot in Q_3 . They are then excited via resonant electric field and the fragments are scanned out, creating the tandem mass spectrum. **d**, The quadrupole TOF instrument combines the front part of a triple quadrupole instrument with a reflector TOF section for measuring the mass of the ions. **e**, The (three-dimensional) ion trap captures the ions as in the case of the linear ion trap, fragments ions of a particular m/z , and then scans out the fragments to generate the tandem mass spectrum. **f**, The FT-MS instrument also traps the ions, but does so with the help of strong magnetic fields. The figure shows the combination of FT-MS with the linear ion trap for efficient isolation, fragmentation and fragment detection in the FT-MS section.

identifying and quantifying the polypeptides as well as tools for integrating and analysing all the data must be used in concert. Out of a bewildering multitude of techniques and instruments, two main tracks can be identified. The first, and most commonly used, is a combination of 2DE and MS. The second track combines limited protein purification with the more recently developed techniques of automated peptide MS/MS and, if accurate quantification is desired, stable-isotope tagging of proteins or peptides. In either track a suitable data processing, storage and visualization infrastructure needs to be developed, if the platform is intended for high-throughput operation.

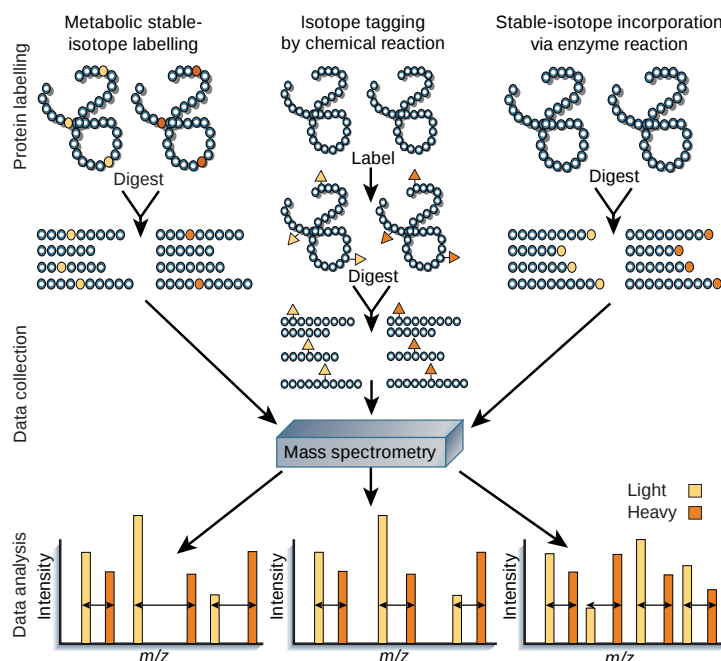
In the first track, the proteins in a sample are separated by 2DE, stained, and each observed protein spot is quantified by its staining intensity. Selected spots are excised, digested and analysed by MS. Sophisticated pattern-matching algorithms as well as interpretation by skilled researchers are required to relate the 2DE patterns to each other in order to detect characteristic patterns and differences among samples. 2DE has been a mature technique for more than 25 years and was the first technique capable of supporting the concurrent quantitative analysis of large numbers of gene products. In fact, many of the principles now commonly used for global, quantitative analysis of messenger RNA expression patterns, such as clustering algorithms and multivariate statistics, were developed in the context of 2DE¹⁸.

Peptide-mass mapping by MALDI-TOF and peptide sequencing by ESI-MS/MS have become highly efficient at the identification of

gel-separated proteins. In the many reports using this technology, largely the same proteins were identified repeatedly, irrespective of the system studied, which suggests limited dynamic range of 2DE-based proteomics. Systematic studies of the budding yeast *Saccharomyces cerevisiae* indeed revealed that typically only the most abundant proteins can be observed by this method¹⁹. Incremental improvements in 2DE technology, including more sensitive staining methods^{20,21}, large-format higher resolving gels²² and sample fractionation prior to 2DE have alleviated, but not eliminated, these and other shortcomings of the 2DE/MS approach.

Studying major histocompatibility complex class I-associated peptides, a natural and complex peptide library, Hunt and colleagues pioneered the use of LC-MS/MS for the analysis of complex peptide mixtures and it is this method that is today at the core of MS-based proteomics²³. However, before LC-MS/MS could be used both for the identification of protein mixtures and for quantitative proteomic experiments, a number of technical issues had to be addressed. First, single-dimension peptide chromatography does not provide sufficient peak capacity to separate peptide mixtures as complex as those generated by the proteolysis of protein mixtures of, for example, total cell lysates. Second, in both MALDI- and ESI-MS, the relationship between the amount of analyte present and measured signal intensity is complex and incompletely understood. Mass spectrometers are therefore inherently poor quantitative devices.

Figure 3 Schematic representation of methods for stable-isotope protein labelling for quantitative proteomics. **a**, Proteins are labelled metabolically by culturing cells in media that are isotopically enriched (for example, containing ^{15}N salts, or ^{13}C -labelled amino acids) or isotopically depleted. **b**, Proteins are labelled at specific sites with isotopically encoded reagents. The reagents can also contain affinity tags, allowing for the selective isolation of the labelled peptides after protein digestion. The use of chemistries of different specificity enables selective tagging of classes of proteins containing specific functional groups. **c**, Proteins are isotopically tagged by means of enzyme-catalysed incorporation of ^{18}O from ^{18}O water during proteolysis. Each peptide generated by the enzymatic reaction carried out in heavy water is labelled at the carboxy terminal. In each case, labelled proteins or peptides are combined, separated and analysed by mass spectrometry and/or tandem mass spectrometry for the purpose of identifying the proteins contained in the sample and determining their relative abundance. The patterns of isotopic mass differences generated by each method are indicated schematically. The mass difference of peptide pairs generated by metabolic labelling is dependent on the amino acid composition of the peptide and is therefore variable. The mass difference generated by enzymatic ^{18}O incorporation is either 4 Da or 2 Da, making quantitation difficult. The mass difference generated by chemical tagging is one or multiple times the mass difference encoded in the reagent used.



Third, the amount of data collected by the method is huge and its analysis daunting.

Substantial progress has been achieved in each of these areas, resulting in the emergence of increasingly robust and productive platforms. To provide more peak capacity, various combinations of protein and peptide separation schemes have been explored. Most popular at present are two-dimensional (strong cation exchange/reversed phase)^{24,25} or three-dimensional (strong cation exchange/avidin/reversed phase)²⁶ chromatographic separations of peptide mixtures generated by tryptic digestion of protein samples that are frequently pre-fractionated by 1DE. Several studies suggest that, in principle, these methods are capable of detecting proteins of very low abundance, although considerable effort is required and a sufficient amount of starting protein sample must be available^{27,28}. However, no proteome has yet been completely analysed and, for lack of a suitable reference, it will be difficult to determine when that milestone has been achieved.

To add a quantitative dimension to peptide LC-MS/MS experiments, the proven technique of stable-isotope dilution has been applied. This method makes use of the facts that pairs of chemically identical analytes of different stable-isotope composition can be differentiated in a mass spectrometer owing to their mass difference, and that the ratio of signal intensities for such analyte pairs accurately indicates the abundance ratio for the two analytes (Fig. 3). To this end, stable-isotope tags have been introduced to proteins via metabolic labelling using heavy salts or amino acids²⁹, enzymatically via transfer of ^{18}O from water to peptides^{30,31}, or via chemical reactions using isotope-coded affinity tags or similar reagents^{32,33}. Post-isolation chemical isotope tagging of proteins is currently the most versatile and most commonly used labelling method. An attractive feature of this approach is that the selectivity of the labelling reactions can be used to direct the isotopes and attached affinity tags to specific functional groups or protein classes, thus enabling their selective isolation and analysis.

So far, isotope-tagging chemistries have been described that are specific for sulphhydryl groups^{32,33}, amino groups³⁴, the active sites for serine³⁵ and cysteine hydrolases³⁶, for phosphate ester groups^{37,38} and for *N*-linked carbohydrates³⁹. Site-specific isotope tagging is limited

only by the creativity of the chemist synthesizing suitable reagents. We therefore expect that new reagents will make many different types of 'sub-proteomes' accessible to quantitative analysis. Recently, a method called 'stable-isotope labelling with amino acids in cell culture', or SILAC, has been described⁴⁰. In this method, one cell state is metabolically labelled by, for example, ^{13}C -labelled arginine. Potentially all peptides can be labelled and the absence of any chemical steps make the method easy to apply as well as compatible with multistage purification procedures.

A current challenge for high-throughput proteomics is to use CID database search results from large numbers of peptide CID spectra to derive a list of identified peptides and their corresponding proteins. This task entails distinguishing correct peptide assignments from false identifications among database search results. In the case of small data sets, this can be achieved by researchers with expertise in spectral interpretation, manually verifying the peptide assignments to spectra made by database search programs. Such a time-consuming approach is not feasible for high-throughput analysis of large data sets containing tens of thousands of spectra, or when expertise is not available.

Alternatively, researchers can attempt to separate the correct from incorrect peptide assignments by applying filtering criteria based upon database search scores and other available data^{25,26,28}. However, the rates of false identifications that result from such filters are not known, nor is it known how those rates are affected by mass spectrometer, sample preparation, or spectrum quality. In addition, researchers often use their own preferred filtering criteria, making it particularly difficult to compare their results among or even within groups. Consequently, the question of what constitutes an identified protein in a LC-MS/MS experiment has been difficult to answer. It is therefore important that computer programs that use robust and transparent statistical principles to estimate accurate probabilities indicating the likelihood for the presence of a peptide or protein in the sample^{41,42} are further developed and widely tested and applied.

The technologies and tools described here are now being combined to create robust platforms for quantitative, high-throughput proteomics. This effort is aided by the introduction of the new types of high-performance mass spectrometers discussed

above. Currently, specialized MS laboratories can easily identify and quantify hundreds of proteins per day on a single MS system, and rapid advances in sample throughput, sensitivity and accuracy are projected.

Applying proteomics technology to protein profiling

Protein mixtures of considerable complexity can now be routinely characterized in some depth using the methods described above. One measure of technical progress is the number of proteins identified in each study. Such numbers can now reach into the thousands for suitably complex samples. But to be biologically useful, as opposed to simply highlighting analytical features of the methods, large-scale proteomic studies need to solve biological questions. In this regard, MS-based proteomics has interfaced particularly well with three types of biological or clinical questions. The first is the generation of protein–protein linkage maps. The second is the use of protein identification technology to annotate and, if necessary, correct genomic DNA sequences. The third is the use of quantitative methods to analyse protein expression profiles as a function of cellular state as an aid to infer cellular function.

The sequences of many mature proteins in higher eukaryotes, after processing and splicing, are often not directly apparent from their cognate DNA sequences. Peptide sequence data of sufficient quality provides unambiguous evidence of translation of a particular gene and can, in principle, differentiate between alternatively spliced or translated forms of a protein. Using a combination of MS and gene chip analysis, a number of proteins that were derived from previously undetected open-reading frames were found in the yeast genome⁴³ and previously unknown human genes have also been found by direct searching of the human genome sequence^{44,45}.

Thus, it might be tempting to systematically analyse the proteins expressed by a cell or tissue, that is, to generate comprehensive proteome maps. First-generation large-scale proteome maps of microorganisms such as yeast⁴⁸ or the bacterium *Deinococcus radiodurans*¹¹ are examples of such projects and, with products from more than 60% of the genes identified, the *Deinococcus* map is at present the most complete. A recent review⁴⁶ of the proteomics of human plasma highlights a number of challenges facing comprehensive blood-serum analysis and thus, by implication, other samples from higher eukaryotes. Considering the combinatorial effects of splicing, processing and PTMs, plasma is estimated to contain many thousands to perhaps millions of polypeptide species, spanning a concentration range of up to 10 orders of magnitude. The fact that only about 500 proteins have so far been reported⁴⁷, and very few have been quantified, illustrates the need for further technological developments to address these issues.

The more common and versatile use of large-scale MS-based proteomics has been to document the expression of proteins as a function of cell or tissue state. We argue that to be meaningful, such data must be at least semi-quantitative and that a simple list of proteins detected in the different states is insufficient. This is because analyses of complex mixtures are often not comprehensive, and therefore the non-appearance of a particular sequence in the list of identified peptides does not indicate that the peptide or protein was not originally present in the sample. Additionally, it is often impossible to prepare a certain cell type, cell fraction or tissue in completely pure form, without trace contaminations of other fractions. And because the ion current of a peptide is dependent on a multitude of variables that are difficult to control, this measure is not a good indicator of peptide abundance. If stable-isotope dilution has not been used, a rough relative estimate of the quantity of the protein can be gained by integrating the ion current of its peptide-mass peaks over their elution time and comparing these 'extracted ion currents' between states, provided that highly accurate and reproducible methods are used.

The malaria parasite *Plasmodium falciparum* has recently been subjected to detailed proteomic analysis. The life cycle of the parasite

is complex; thus there is great interest in the proteins it expresses in its own different stages and in its different host compartments. Illustrating the power and importance of proteomics, the recent genome project of the malaria parasite was accompanied by two large-scale proteome efforts. In one of the studies⁴⁸ the human stages of the parasite were analysed and a large number of proteins identified in the sexual and non-sexual stages. Quantitation was attempted by comparing peptide ion currents between stages and by correlation of protein data with RNA quantification by the polymerase chain reaction. After bioinformatic analysis, a set of proteins was selected from membrane fractions for follow-up as possible stage-specific drug or vaccine targets. The study resulted in more than 200 such candidate proteins and generated a large set of 'orphan' peptides that were not found in the set of predicted proteins, but were mapped onto the genome, assisting its annotation.

The other *P. falciparum* proteomics project analysed mosquito and human stages of the parasite and reported a total of around 2,400 identified proteins⁴⁹. The study revealed unexpected stage specificity of a number of surface proteins and suggested co-expression of new proteins with groups of proteins already annotated as stage specific, helping to place these proteins in a functional context. The study also illustrated the need for transparent statistical tools to improve the confidence in protein identifications, as a large number, and in some cases the majority, of the proteins were identified solely by single peptides, many of which did not conform to the expected tryptic cleavage pattern.

Increasingly, stable-isotope dilution and LC-MS/MS are used to accurately detect changes in quantitative protein profiles and to infer biological function from the observed patterns. Shiio *et al.*⁵⁰ identified the reduction of stress fibres and focal adhesions as a new cellular function of the Myc oncogene by comparing protein extracts from Myc⁺ and Myc⁻ cells. Han *et al.*²⁶ identified pleiotropic, differentiation-induced effects in the microsomal compartment of phorbol ester-treated HL-60 cells, and a number of studies have also identified previously unknown connections between metabolic processes^{51,52}.

Applying proteomics technology to protein interactions

The analysis of protein complexes is the third area where MS-based proteomics has had a significant impact. Most proteins exert their function by way of protein–protein interactions and enzymes are often held in tightly controlled regions of the cell by such interactions. Thus, one of the first questions usually asked about a new protein — apart from where it is expressed — is to what proteins does it bind? To study this question by MS, the protein itself is used as an affinity reagent to isolate its binding partners. Compared with two-hybrid and chip-based approaches, this strategy has the advantages that the fully processed and modified protein can serve as the bait, that the interactions take place in the native environment and cellular location, and that multicomponent complexes can be isolated and analysed in a single operation⁵³. However, because many biologically relevant interactions are of low affinity, transient and generally dependent on the specific cellular environment in which they occur, MS-based methods in a straightforward affinity experiment will detect only a subset of the protein interactions that actually occur. Bioinformatics methods, correlation of MS data with those obtained by other methods, or iterative MS measurements possibly in conjunction with chemical crosslinking⁵⁴ can often help to further elucidate direct interactions and overall topology of multiprotein complexes.

MS-based protein interaction experiments have three essential components: bait presentation, affinity purification of the complex, and analysis of the bound proteins. Ideally, endogenous proteins can serve as bait if an antibody or other reagent exists that allows specific isolation of the protein with its bound partners. Unfortunately, there are currently no comprehensive antibody collections and many current antibodies do not immunoprecipitate well or lack sufficient specificity. A more generic strategy is to 'tag' the proteins of interest with a sequence readily recognized by an antibody specific for the tag.

To facilitate expression of the tagged protein at close to physiological levels, the tagged construct is preferably expressed from the promoter of its native, untagged counterpart. This can be achieved in a limited number of species, most notably *S. cerevisiae*, by using homologous recombination to replace the endogenous gene in the genome with a gene coding the tagged protein.

In mammalian cells, where expression of tagged proteins from the native promoter is more difficult, they are usually expressed after transient transfection, in stable cell lines generated by traditional selection, or by recently introduced kits for fast generation of stable cell lines. Transient or stable transfections usually result in tagged-protein expression levels that are different from the untagged, endogenous counterpart. They are therefore prone to artefacts generated by non-physiological levels of the bait protein. Considerable efforts have been devoted to developing tagging systems optimized for analysis of protein complexes (see review in this issue by Fields and co-workers, page 208). Tags supporting single-step purification have the advantage of convenience and yield. Tags supporting two sequential affinity steps (tandem affinity purification or TAP) combine two different tags on the same protein, which are normally separated by an enzyme-cleavable linker sequence.

A popular implementation of this concept consists of a calmodulin-binding domain in series with the immunoglobulin-binding domain of protein A, the domains being separated by a sequence that can be cleaved by a tobacco etch virus (TEV) protease⁵⁵. The tagged proteins are bound initially to a solid support modified with immunoglobulins, recovered by TEV proteolysis and bound to a calmodulin column from which they can be selectively eluted by increased $[Ca^{2+}]$. TAP tags significantly reduce background noise, but probably result in the loss of some of the more transient and weak binding partners during the purification procedure, as the second affinity step essentially causes infinite sample dilution. The identification part of the strategy is similar to the generic protein identification experiment described above, and essentially all the strategies discussed here have been used for the analysis of protein complexes. However, it is clear that the cleaner the initial purification, the less challenging the mass spectrometric 'readout' becomes.

Combining such developments, two large-scale projects have recently been reported on the protein–protein interaction network in yeast. In one of the studies, 1,739 TAP-tagged genes were introduced into the yeast genome by homologous recombination, 232 stable complexes were isolated and protein constituents were identified by MALDI peptide mapping after separation by 1DE⁵⁶. Apart from the large number of new interactions for known and new yeast proteins, a higher-order interaction structure between complexes emerged from the data. A similar study used transient transfection to express FLAG-tagged bait proteins; complexes were isolated by single-step immunoprecipitation and any attached proteins identified by automated LC-MS/MS of gel-separated bands⁵⁷. These experiments probed the phosphatases, kinases and the DNA-repair network of yeast specifically, resulting in many interesting signalling connections being made.

Both studies reported a large number of interacting proteins, and while groups of selected bait proteins were only partly overlapping, a number of interesting conclusions could be drawn from a comparison of the results. First, protein complexes isolated in a single step resulted in more complex samples than those isolated by the TAP tag procedure. Second, surprisingly little overlap of the data was observed when results from similar bait proteins were compared⁵⁸ between the two studies or between the MS and previous yeast two-hybrid studies. Although a variety of technical explanations have been advanced to explain this discrepancy, it is important to note that the 'interactome' is potentially very large, growing with the square of the number of proteins involved, and that it remains substantially undersampled. Third, both projects reported results consistent with previous literature for already known complexes. As in other large-scale projects, higher accuracy can be obtained with more detailed

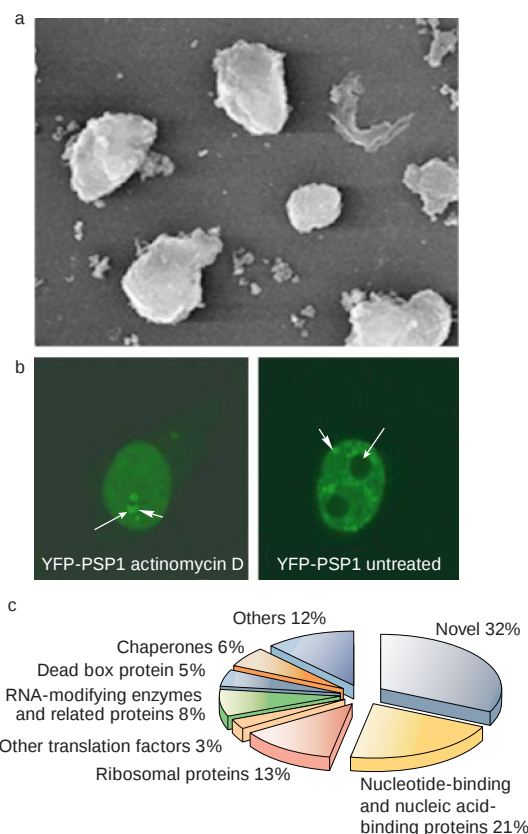


Figure 4 Organellar proteomics by combined mass spectrometry (MS) and imaging methods. An organelle is isolated by biochemical methods such as density-gradient centrifugation. **a**, Scanning electron microscopy shows the result of purification of the human nucleolus. The organellar proteins are then solubilized, optionally fractionated, and analysed by the methods described in Fig. 1; this results in a list of proteins that are candidate constituents of the complex. At this stage MS-based proteomics is combined with imaging techniques such as laser confocal scanning fluorescence microscopy of yellow fluorescent protein (YFP)-tagged members of the complex. **b**, The right panel shows localization of one of the new proteins found by proteomics to subnuclear structures termed 'paraspeckles'. The left panel shows that inhibition of transcription by actinomycin D leads to reorganization of the nucleoli and concentration of the novel protein to cap structures⁹¹. **c**, Functional categorization of the proteins in the organelle. More than 400 proteins have been found and can be assayed for dynamic localization and biochemical function in the context of the nucleolus.

experiments, for example, in a 'complex walking' strategy in which complexes are tagged and identified sequentially⁵⁹.

In the future, quantitative methods based on stable-isotope labelling are likely to revolutionize the study of stable or transient interactions and interactions dependent on PTMs. In such experiments, accurate quantification by means of stable-isotope labelling is not used for protein quantification per se; instead the stable-isotope ratios distinguish between the protein composition of two or more protein complexes. In the case of a sample containing a complex and a control sample containing only contaminating proteins (for example, immunoprecipitation with an irrelevant antibody or isolate from a cell devoid of affinity-tagged protein), the method can distinguish between true complex components and nonspecifically associated proteins. In the case of complexes isolated from cells at different states (for example, activated and non-activated cells) the method can identify dynamic changes in the composition of protein complexes^{60,61}.

The ability of quantitative MS to detect specific complex components within a background of nonspecifically associated proteins increases the tolerance for high background and allows for fewer purification steps and less stringent washing conditions, thus increasing the chance of finding transient and weak interactions. The same methods can be used to study the interaction of proteins with nucleic acids, small molecules and in fact with any other substrate. For example, drugs can be used as affinity baits in the same way as proteins to define their cellular targets, and small molecules such as co-factors can be used to isolate interesting 'sub-proteomes'⁶².

MS-based proteomics is not limited to the analysis of complexes consisting of only a few proteins. In fact, some of the most biologically informative results have come from the analysis of large protein complexes — 'molecular machines', organelles and subcompartments of the cell. The first complex analysed in this way was the spliceosome, studied in yeast⁶³ and then in human cells⁶⁴, closely followed by the yeast nuclear pore complex⁶⁵. Re-analysis of the spliceosome using more complete databases and more advanced instrumentation has recently been undertaken. In one study, nearly 300 proteins were found, and evidence from sequence analysis highlighted a set of 55 novel proteins involved in splicing and RNA processing⁶⁶. A similar study, using an elegant RNA tag-based purification, also discovered many new proteins⁶⁷. Both studies found essentially the complete list of known human splicing factors. The new data encompassed and extended the original results, indicating the maturity of MS-based methods for the analysis of such complex structures. The next challenge will now be to study the dynamics and assembly of functional protein modules via quantitative proteomics.

Numerous other large complexes and organelles have now at least been partly characterized by MS⁶⁸. The limiting factor in such experiments is no longer primarily the analysis, but rather the ability to purify such structures to homogeneity. For example, it is very difficult to isolate structures such as the Golgi apparatus and the interpretation of results from samples of dubious quality and definition is correspondingly vague. The largest organelle mapped so far is the human nucleolus, whose high specific density allows for a simple, efficient purification⁴⁵ (Fig. 4). By using a variety of mass spectrometric techniques, more than 400 nucleolar proteins have now been identified in this structure. Well-characterized proteins identified in this study, but not previously known to be associated with nucleolus, raise interesting questions about the function of this organelle, while the identification of a large number of previously uncharacterized gene products places many of those in the context of nucleolar function.

At the same time, some of the previously known nucleolar proteins, such as Werner's syndrome protein, have not yet been found, indicating that even this large-scale study is not yet complete. One reason for this is that numerous factors, including Werner's syndrome protein, exhibit either cell-cycle dependent or facultative interactions with nucleoli. Dynamic imaging studies of the nucleus also make it clear that many of the factors in the nucleolus are associated only transiently with this organelle⁶⁹, a fact reflected in the overlapping 'cast of characters' of several nuclear bodies studied. Just as the single protein/single function concept is turning out to be more the exception than the rule, the concept of a single subcellular location of a protein may also turn out to be a gross over-simplification.

Applying proteomics to the analysis of protein modifications

Proteins are converted to their mature form through a complicated sequence of post-translational protein processing and 'decoration' events. Many of the PTMs are regulatory and reversible, most notably protein phosphorylation, which controls biological function through a multitude of mechanisms. Mass spectrometric methods to determine the type and site of such modifications on single, purified proteins have been refined over the past two decades. In this case, peptide mapping with different enzymes is usually used to 'cover' as much of the protein sequence as possible. Protein modifications are then determined by examining the measured mass and fragmenta-

tion spectra via manual or computer-assisted interpretation. For the analysis of some types of PTMs, specific mass spectrometric techniques have been developed that scan the peptides derived from a protein for the presence of a particular modification. The analysis of regulatory modifications, in particular protein phosphorylation, is complicated by the frequently low stoichiometry, the size and ionizability of peptides bearing the modifications, and their fragmentation behaviour in the mass spectrometer (reviewed in refs 4,70,71). The analysis of the modification state of a purified protein therefore remains a challenging analytical endeavour.

Recently, attempts have been made to define modifications on a proteome-wide scale. Given the difficulties of identifying all modifications even in a single protein, it is clear that, at present, scanning for proteome-wide modifications is not comprehensive. Nevertheless, a large amount of biologically useful information can, in principle, be generated by this approach. One of the strategies used is essentially an extension of the approach used to analyse protein mixtures⁷². Instead of searching the database only for non-modified peptides, the database search algorithm is instructed to also match potentially modified peptides. To avoid a 'combinatorial explosion' resulting from the need to consider all possible modifications for all peptides in the database, the experiment is usually divided into identification of a set of proteins on the basis of non-modified peptides, followed by searching only these proteins for modified peptides⁷².

A more functionally oriented strategy focuses on the search for one type of modification on all the proteins present in a sample. Such techniques are based usually on some form of affinity selection that is specific for the modification of interest and which is used to purify the 'sub-proteome' bearing this modification. For example, Pandey *et al.* stimulated cells with epidermal growth factor and isolated newly phosphotyrosine-modified proteins using antibodies specific to phosphotyrosine^{73,74}. Efforts to determine the 'phosphoproteome' in a single step^{37,38} have used chemical modification combined with affinity selection. In a potentially powerful technique, Ficcaro *et al.*⁷⁵ esterified peptide mixtures, thereby nullifying negatively charged carboxyl groups, and then captured phosphopeptides on metal affinity columns. This approach overcomes the low specificity of these columns caused by their affinity for any negatively charged peptides and seems to significantly improve the capture of phosphopeptides. Further development of these and related techniques may allow study of the complete phosphoproteome in multiprotein complexes and the pattern of the more abundant phosphopeptides in whole cells, a promising approach to study the activation state of whole signalling networks.

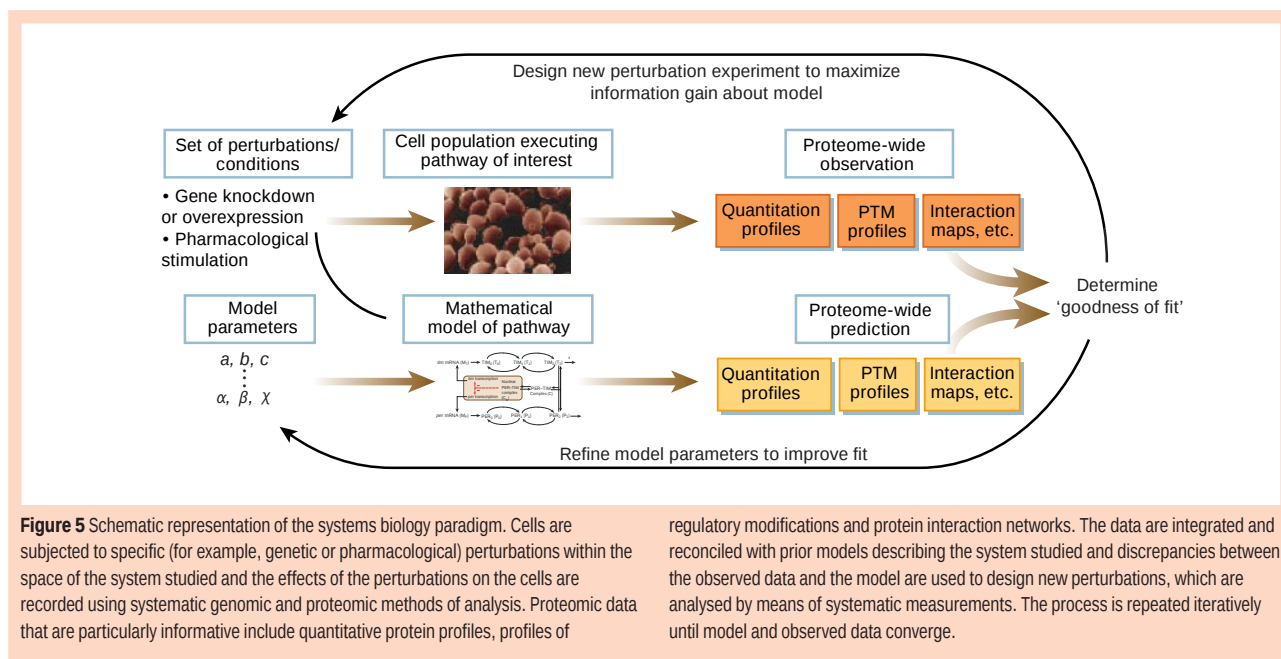
Gygi *et al.* have used affinity purification to capture the ubiquitinated proteins of yeast cells (ref. 76 and S. P. Gygi, personal communication). Over 1,000 such proteins were identified and in more than 100 cases the site of ubiquitination was determined. These results open up the study of ubiquitinated substrates in a cell state- and protein complex-dependent manner.

Many challenges remain in the large-scale mapping of PTMs, but it is clear that MS-based proteomics can make a unique contribution in this area. For example, systematic quantitative measurements of PTMs by stable-isotope labelling would be of tremendous biological interest.

Challenges, expectations and emerging technologies

Proteomics, in particular quantitative proteomics, can be viewed as an array of biological or clinical assays capable of probing most, if not all, of the proteins in a sample. As proteins are involved in essentially all biological functions and clinical conditions, MS and proteomics will have an even greater impact on biology and medicine than it has had so far.

Over the past decade, MS of single proteins or protein complexes has been successful to the point where it is now considered a mainstream technology. This technology interfaces particularly well with biochemical and cell biological studies for studying specific protein functions. The success is built on the proven potential of mass spectrometric techniques to rapidly identify almost any protein, to



analyse that protein for the presence of PTMs, to determine how and with what other biomolecules that proteins interacts, and even to gain structural information about the protein from gas-phase experiments^{77,78} and from experiments in which mass spectrometric characterization of proteins has interfaced with X-ray crystallography⁷⁹. As analytical methods and instrumentation are improving constantly, MS can be used to address an increasing number of analytical problems facing biochemists, geneticists and cell biologists. But protein MS does not equal proteomics. The specific objective of proteomics is to concurrently identify, quantify and analyse a large number of proteins in a functional context. This shift in focus from the analysis of selected isolated proteins to proteome-wide analyses has a number of profound implications and poses as yet unmet challenges for every aspect of experimental biology. These include experimental design, data analysis, visualization and storage, organization of proteomics research groups and publication of proteomic data.

Experimental design

In a typical protein MS experiment, a specific property (for example, sequence, PTM or interaction) of a partly characterized protein is examined. In contrast, proteomic experiments often collect large amounts of data in the absence of hypotheses concerning specific proteins or activities. Proteomic experiments, therefore, have to be designed in ways that maximize the likelihood of generating new discoveries, or at least new testable hypotheses. The technology of gene expression profiling is conceptually similar to proteomic profiling and has demonstrated that more information is better. Although it is essentially impossible to draw meaningful conclusions from a single quantitative gene expression profile, the availability of multiple profiles from related samples allows the application of statistical tools⁸⁰ to extract signature patterns containing diagnostic or functional information. Therefore, successful proteomics experiments need to be designed in such a way that they can take advantage of the power of statistics for data interpretation. To achieve this goal, carefully controlled repeat studies and the generation of models describing the source, magnitude and distribution of errors will be essential.

Data collection

Proteomic studies necessarily result in large amounts of data. Data collection at a volume and quality that is consistent with the use of

statistical methods is a significant limitation of proteomics today. In a typical LC-MS/MS experiment, approximately 1,000 CID spectra can be acquired per hour. Even with the optimistic assumption that every one of these spectra leads to the successful identification of a peptide, it would take a long time to analyse complete proteomes. High-throughput collection of consistently high-quality data therefore remains a challenge in proteomics. We have argued that one solution to the problem would be to establish a number of specialized and generally accessible data-collection centres⁸¹, akin to the beam lines used by X-ray crystallographers for protein structural studies. Such centres would not only generate data of consistent quality for a large number of proteomics projects, but would also serve as disseminators of advanced technology.

Data analysis, visualization and storage

The analysis and interpretation of the enormous volumes of proteomic data remains an unsolved challenge, particularly for gel-free approaches. Expert manual analysis is incompatible with the tens of thousands of spectra collected in a single experiment and is inconsistent. Therefore, the development of transparent tools for the analysis of proteomic data using statistical principles is a key challenge^{41,42}. Only once such tools are tested, validated and widely accepted will it become feasible to apply quality standards for protein identification, quantification and other measurements and to compare complementary proteomic data sets generated in different laboratories. These comparisons will also depend critically on transparent file structures for data storage, communication and visualization. The development of such proteomics tools is still in its infancy.

Data publication

The publication of the large data sets generated by proteomic experiments and the information contained therein poses significant challenges. At present, most proteomics publications consist of an experimental description, a data table (typically published as supplementary material containing a partially interpreted and validated summary of the data) and an in-depth validation and discussion of one to a few conclusions made from the data. To make publication of proteomics data more useful, publishers and journals need to find new ways to review large data sets, to validate their contents and to make the information contained therein electronically searchable;

this problem remains essentially unsolved, despite preliminary developments by a few publishers and journals⁸².

In spite of these and other challenges, the impact of proteomics on clinical and biological research is growing rapidly. It seems that beyond its great current contribution to cell biology, proteomics may have a huge influence on clinical diagnosis. MS-based proteomics seems capable of detecting patterns of differentially expressed proteins in easily accessible clinical samples such as blood serum. These types of analyses have the potential to diagnose the presence and stage of many diseases, in particular cancers⁸³. Clinical diagnosis will be further advanced with the advent of mass spectrometers with higher mass accuracy, dynamic range and resolution, and with the ability to identify specific sequences of diagnostic analytes and the use of accurate quantification procedures.

MS-based proteomics is still an emerging technology where revolutionary change is possible. Several concepts have been proposed and are under development that have the potential to alter the landscape of current MS-based proteomic technologies. One of these is the analysis of intact proteins. The currency of essentially all MS-based identifications is peptides. The convergence of mass spectrometers with large mass ranges, extremely high mass accuracy and resolution, and ionization/fragmentation methods compatible with large proteins has catalysed the emergence of whole-protein proteomics⁸⁴. The analysis of whole proteins with high accuracy has the potential to distinguish and characterize differentially modified forms and to provide insights into coordinated modification patterns that are difficult to establish by peptide analysis.

A second emerging concept is mass spectrometric tissue imaging⁸⁵. In this technique, thin tissue sections are directly applied to a MALDI mass spectrometer and, after treating the samples with a suitable matrix, profiles of the proteins contained in the section are generated by 'imaging' the sample with an array of mass spectra. The method, while currently incapable of identifying the detected protein features, has already provided proof-of-principle that clinically diagnostic patterns can be generated. Increased spatial resolution, potentially to subcellular levels, improved software tools and automated sample preparation will further increase the utility of this technique for clinical diagnosis and classification.

A third concept is the use of mass tags measured in mass spectrometers of very high mass accuracy and resolution such as FT-MS instruments. These mass tags could be used potentially for high-throughput protein identification. The idea is based on the observation that a particular proteome, if digested with a specific enzyme such as trypsin, will generate a peptide mixture in which most peptides can be uniquely classified based on their accurate mass and some other parameters such as chromatographic coordinates^{86,87}. Therefore, once the peptides are identified by MS/MS and annotated with accurate mass tags they can be identified in subsequent experiments simply by correlating the accurate mass and the separation coordinates with the list of previously determined mass tags.

Conclusion and perspective

In studying a biological system using the biochemical approach, researchers have traditionally attempted to purify to homogeneity each of the system's components; each element is then studied in detail with the ultimate aim being to reconstitute the system *in vitro* from the isolated components. Because proteins carry out most biological activities, the biochemical approach has been significantly enhanced by the availability of the sensitive and rapid MS-based protein identification methods discussed in this article. The availability of complete genomic sequences from a number of species further facilitates MS-based protein identifications, as the requirement for *de novo* sequencing has been usurped by simple correlation of measured data versus theoretical data predicted from sequence databases. The availability of completely sequenced genomes also catalysed the emergence of systems biology — the attempt to system-

atically study all the concurrent physiological processes in a cell or tissue by global measurement of differentially perturbed states (Fig. 5). The ultimate goal of systems biology is the integration of data from these observations into models that might, eventually, represent and simulate the physiology of the cell⁸⁸.

Proteomics is an essential component of systems biology research because proteins are rich in information that has turned out to be extremely valuable for the description of biological processes. These include protein abundances, linkage maps to other proteins or to other types of biomolecules including DNA and lipids, activities, modification states, subcellular location and more. Unfortunately, with the exception of quantitative protein profiles and protein-protein interactions (keeping in mind the caveats discussed above), none of these properties can currently be measured systematically, quantitatively and with high throughput. But rapid advances in technology suggest that this limitation may be transient. The few studies where the same biological system was subjected to different types of systematic measurements already offer insights into the power of the method. For instance, mRNA expression profiles and protein expression profiles seem to be largely complementary and therefore contribute to a more refined description of the system that each observation by itself is unable to provide⁸⁸.

Extrapolating from these limited studies, we expect that combining different genomic and proteomic results obtained from the same biological system will substantially increase our understanding of complex biological processes. More specifically, the systems biology studies based on diverse and high-quality proteomic data will define functional biological modules, reveal previously unrecognized connections between biochemical processes and modules, and generate new hypotheses that can be tested either by traditional methods or by the targeted generation of more genomic and proteomic data^{51,88–90}. □

doi:10.1038/nature01511

- Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
- Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for the mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).
- Karas, M. & Hillenkamp, F. Laser desorption/ionization of proteins with molecular mass exceeding 10000 daltons. *Anal. Chem.* **60**, 2299–2301 (1988).
- Aebersold, R. & Goodlett, D. R. Mass spectrometry in proteomics. *Chem. Rev.* **101**, 269–295 (2001).
- Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437–473 (2001).
- Hager, J. W. A new linear ion trap mass spectrometer. *Rapid Commun. Mass Spectrom.* **16**, 512–526 (2002).
- Schwartz, J. C., Senko, M. W. & Syka, J. E. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **13**, 659–669 (2002).
- Marshall, A. G., Hendrickson, C. L. & Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* **17**, 1–35 (1998).
- Valaskovic, G. A., Kelleher, N. L. & McLafferty, F. W. Atomole protein characterization by capillary electrophoresis-mass spectrometry. *Science* **273**, 1199–1202 (1996).
- Martin, S. E., Shabanowitz, J., Hunt, D. F. & Marto, J. A. Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **72**, 4266–4274 (2000).
- Lipton, M. S. *et al.* Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl Acad. Sci. USA* **99**, 11049–11054 (2002).
- Krutchinsky, A. N., Kalkum, M. & Chait, B. T. Automatic identification of proteins with a MALDI-quadrupole ion trap mass spectrometer. *Anal. Chem.* **73**, 5066–5077 (2001).
- Medzhidzadzky, K. F. *et al.* The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal. Chem.* **72**, 552–558 (2000).
- Loboda, A. V., Krutchinsky, A. N., Brominski, M., Ens, W. & Standing, K. G. A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption/ionization source: design and performance. *Rapid Commun. Mass Spectrom.* **14**, 1047–1057 (2000).
- Mann, M. & Wilm, M. S. Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399 (1994).
- Eng, J. K., McCormack, A. L. & Yates, J. R. I An approach to correlate MS/MS data to amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
- Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
- Anderson, N. L., Hofmann, J. P., Gemmell, A. & Taylor, J. Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clin. Chem.* **30**, 2031–2036 (1984).
- Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl Acad. Sci. USA* **97**, 9390–9395 (2000).
- Rabilloud, T. Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* **2**, 3–10 (2002).
- Unlu, M., Morgan, M. E. & Minden, J. S. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **18**, 2071–2077 (1997).

22. Gauss, C., Kalkum, M., Lowe, M., Lehrach, H. & Klose, J. Analysis of the mouse proteome. (I) Brain proteins: separation by two-dimensional electrophoresis and identification by mass spectrometry and genetic variation. *Electrophoresis* **20**, 575–600 (1999).
23. Hunt, D. F. *et al.* Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263 (1992).
24. Wolters, D. A., Washburn, M. P. & Yates, J. R. III An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).
25. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17**, 676–682 (1999).
26. Han, D. K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnol.* **19**, 946–951 (2001).
27. Gygi, S. P., Rist, B., Griffin, T. J., Eng, J. & Aebersold, R. Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags. *J. Proteome Res.* **1**, 47–54 (2002).
28. Washburn, M. P., Wolters, D. & Yates, J. R. III Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19**, 242–247 (2001).
29. Conrads, T. P., Issaq, H. J. & Veenstra, T. D. New tools for quantitative phosphoproteome analysis. *Biochem. Biophys. Res. Commun.* **290**, 885–890 (2002).
30. Mirgorodskaya, O. A. *et al.* Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using ^{18}O -labeled internal standards. *Rapid Commun. Mass Spectrom.* **14**, 1226–1232 (2000).
31. Yao, X., Freas, A., Ramirez, J., Demirev, P. A. & Fenselau, C. Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem.* **73**, 2836–2842 (2001).
32. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**, 994–999 (1999).
33. Zhou, H., Ranish, J. A., Watts, J. D. & Aebersold, R. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nature Biotechnol.* **20**, 512–515 (2002).
34. Munchbach, M., Quadroni, M., Miotto, G. & James, P. Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal. Chem.* **72**, 4047–4057 (2000).
35. Liu, Y., Patricelli, M. P. & Cravatt, B. F. Activity-based protein profiling: the serine hydrolases. *Proc. Natl Acad. Sci. USA* **96**, 14694–14699 (1999).
36. Greenbaum, D., Medzihradsky, K. F., Burlingame, A. & Bogoy, M. Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chem. Biol.* **7**, 569–581 (2000).
37. Zhou, H., Watts, J. D. & Aebersold, R. A systematic approach to the analysis of protein phosphorylation. *Nature Biotechnol.* **19**, 375–378 (2001).
38. Oda, Y., Nagasu, T. & Chait, B. T. Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nature Biotechnol.* **19**, 379–382 (2001).
39. Zhang, H., Li, X.-J., Martin, D. & Aebersold, R. Quantitative analysis of glycoproteins: applications to serum and membrane proteins. (submitted).
40. Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
41. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
42. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* DOI: 10.1021/pr025556v (2002).
43. Oshiro, G. *et al.* Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res.* **12**, 1210–1220 (2002).
44. Kuster, B., Mortensen, P., Andersen, J. S. & Mann, M. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1**, 641–650 (2001).
45. Andersen, J. S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12**, 1–11 (2002).
46. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
47. Adkins, J. N. *et al.* Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteomics* DOI: 10.1074/mcp.M200066-MCP200 (2002).
48. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537–542 (2002).
49. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).
50. Shiio, Y. *et al.* Quantitative proteomic analysis of Myc oncoprotein function. *EMBO J.* **21**, 5088–5096 (2002).
51. Griffin, T. J. *et al.* Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **1**, 323–333 (2002).
52. Baliga, N. S. *et al.* Coordinate regulation of energy transduction modules in *Halobacterium* sp. analyzed by a global systems approach. *Proc. Natl Acad. Sci. USA* **99**, 14913–14918 (2002).
53. Ashman, K., Moran, M. F., Sicheri, F., Pawson, T. & Tyers, M. Cell signalling—the proteomics of it all. *Science's STKE* <http://stke.sciencemag.org/cgi/content/full/sigtrans;2001/103/pe33> (2001).
54. Rappsilber, J., Siniosoglou, S., Hurt, E. C. & Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.* **72**, 267–275 (2000).
55. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17**, 1030–1032 (1999).
56. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
57. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
58. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
59. Shevchenko, A., Schaft, D., Roguev, A., Pijnappel, W. W. & Stewart, A. F. Deciphering protein complexes and protein interaction networks by tandem affinity purification and mass spectrometry: analytical perspective. *Mol. Cell. Proteomics* **1**, 204–212 (2002).
60. Blagoev, B. *et al.* A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nature Biotechnol.* advance online publication, 10 February 2003 (doi:10.1038/nbt790).
61. Ranish, J. A. *et al.* The study of macromolecular complexes by quantitative proteomics. *Nature Genet.* (in the press).
62. MacDonald, J. A., Mackey, A. J., Pearson, W. R. & Haystead, T. A. A strategy for the rapid identification of phosphorylation sites in the phosphoproteome. *Mol. Cell. Proteomics* **1**, 314–322 (2002).
63. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl Acad. Sci. USA* **94**, 385–390 (1997).
64. Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20**, 46–50 (1998).
65. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148**, 635–651 (2000).
66. Rappsilber, J., Ryder, U., Lamond, A. I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12**, 1231–1245 (2002).
67. Zhou, Z., Licklider, L. J., Gygi, S. P. & Reed, R. Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182–185 (2002).
68. Taylor, S. W., Fahy, E. & Ghosh, S. S. Global organellar proteomics. *Trends Biotechnol.* **21**, 82–88 (2003).
69. Leung, A. K. & Lamond, A. I. In vivo analysis of NHPX reveals a novel nucleolar localization pathway involving a transient accumulation in splicing speckles. *J. Cell Biol.* **157**, 615–629 (2002).
70. Mann, M. *et al.* Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol.* **20**, 261–268 (2002).
71. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nature Biotechnol.* (in the press).
72. MacCoss, M. J. *et al.* Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl Acad. Sci. USA* **99**, 7900–7905 (2002).
73. Pandey, A. *et al.* Analysis of receptor signaling pathways by mass spectrometry: identification of Vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc. Natl Acad. Sci. USA* **97**, 179–184 (2000).
74. Steen, H., Kuster, B., Fernandez, M., Pandey, A. & Mann, M. Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway. *J. Biol. Chem.* **277**, 1031–1039 (2002).
75. Ficarro, S. B. *et al.* Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnol.* **20**, 301–305 (2002).
76. Peng, J. & Gygi, S. P. Proteomics: the move to mixtures. *J. Mass Spectrom.* **36**, 1083–1091 (2001).
77. Hanson, C. L., Fucini, P., Ilag, L. L., Nierhaus, K. H. & Robinson, C. V. Dissociation of intact *Escherichia coli* ribosomes in a mass spectrometer—evidence for conformational change in a ribosome elongation factor G complex. *J. Biol. Chem.* **278**, 1259–1267 (2002).
78. Oh, H. *et al.* Secondary and tertiary structures of gaseous protein ions characterized by electron capture dissociation mass spectrometry and photofragment spectroscopy. *Proc. Natl Acad. Sci. USA* **99**, 15863–15868 (2002).
79. Cohen, S. L. & Chait, B. T. Mass spectrometry as a tool for protein crystallography. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 67–85 (2001).
80. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
81. Aebersold, R. & Watts, J. D. The need for national centers for proteomics. *Nature Biotechnol.* **20**, 651 (2002).
82. Mann, M. A home for proteomics data? *Nature* **420**, 21 (2002).
83. Petricoin, E. F. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577 (2002).
84. Mørtz, E. *et al.* Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl Acad. Sci. USA* **93**, 8264–8267 (1996).
85. Stoeckli, M., Chaurand, P., Hallahan, D. E. & Caprioli, R. M. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature Med.* **7**, 493–496 (2001).
86. Goodlett, D. R. *et al.* Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. *Anal. Chem.* **72**, 1112–1118 (2000).
87. Smith, R. D. *et al.* An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2**, 513–523 (2002).
88. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
89. Betts, J. C., Lukey, P. T., Robb, L. C., McAdam, R. A. & Duncan, K. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol. Microbiol.* **43**, 717–731 (2002).
90. Guina, T. *et al.* Quantitative proteomic analysis of *Pseudomonas aeruginosa* indicates synthesis of quinolone signal in adaptation to cystic fibrosis airways. *Proc. Natl Acad. Sci. USA* (in the press).
91. Fox, A. H. *et al.* Paraspeckles. A novel nuclear domain. *Curr. Biol.* **12**, 13–25 (2002).

Acknowledgements We thank members of the Institute of Systems Biology (ISB) and the Center for Experimental BioInformatics (CEBI) for critical reading of the manuscript, preparation of figures and fruitful discussions, especially L. Foster, S.-E. Ong, J. Andersen and L. Feltz. CEBI is supported by a grant from the Danish Natural Research Foundation. R.A. is supported by grants from the National Institute of Health and Oxford GlycoSciences, and a contract from the National Heart, Lung, and Blood Institute, National Institutes of Health. The ISB is supported in part by a gift from Merck and Co.

Protein analysis on a proteomic scale

Eric Phizicky*, Philippe I. H. Bastiaens†, Heng Zhu‡, Michael Snyder‡ & Stanley Fields§

*University of Rochester School of Medicine, Department of Biochemistry and Biophysics, Box 712, 601 Elmwood Avenue, Rochester, New York 14642, USA (e-mail: eric_phizicky@urmc.rochester.edu)

†European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany (e-mail: philippe.bastiaens@embl-heidelberg.de)

‡Department of Molecular, Cellular, and Developmental Biology, PO Box 208103, Yale University, New Haven, Connecticut 06520, USA (e-mail: heng.zhu@yale.edu; michael.snyder@yale.edu)

§Howard Hughes Medical Institute, Departments of Genome Sciences and Medicine, University of Washington, Box 357730, Seattle, Washington 98195, USA (e-mail: fields@u.washington.edu)

The long-term challenge of proteomics is enormous: to define the identities, quantities, structures and functions of complete complements of proteins, and to characterize how these properties vary in different cellular contexts. One critical step in tackling this goal is the generation of sets of clones that express a representative of each protein of a proteome in a useful format, followed by the analysis of these sets on a genome-wide basis. Such studies enable genetic, biochemical and cell biological technologies to be applied on a systematic level, leading to the assignment of biochemical activities, the construction of protein arrays, the identification of interactions, and the localization of proteins within cellular compartments.

Proteomics — the analysis of genomic complements of proteins — has burst onto the scientific scene with stunning rapidity over the past few years, perhaps befitting a discipline that can enjoy the virtually instantaneous conversion of a genome sequence to a set of predicted proteins. But whereas every fragment of DNA behaves biochemically much like any other, proteins possess unique properties, and such individuality creates an enormous hurdle for methodologies that seek to assign an activity to sets of proteins that may number in the thousands¹. Yet the confluence of breakthroughs in cloning and expression technologies, biochemical and genetic strategies, and the instrumentation of mass spectrometry and microscopy has made such global assays increasingly common.

We describe some of these technologies and strategies here, along with a discussion of their advantages and disadvantages, and a brief consideration of new technologies still at the design stage.

Protein expression and purification

The development of methods for parallel analysis of the proteome has relied on the rapid identification of open reading frames (ORFs) and their facile cloning and manipulation. An ORF is defined as the amino acid codons between the initiation codon at the start and the termination codon at the end. ORF identification can be complicated by uncertainties in defining translation start sites, small size and, in particular, the signals for splicing, polyadenylation and editing that can lead to multiple messenger RNA species from a single DNA sequence. Even for a simple and well-studied eukaryote such as the yeast *Saccharomyces cerevisiae*, in which RNA processing is relatively uncomplicated, the number of ORFs has been revised several times as a result of transcriptional analysis and the comparative analysis of genomes of close relatives^{2,3}.

Although cloning of a genomic set of ORFs enables the technologies discussed here to be performed, it is important to note that such a step entails the loss of much of the natural diversity of proteins. For example, a single spliced mRNA is generally chosen as a template for each gene, and the many other mRNA species that result in protein isoforms are not considered. Similarly, post-translational modifications,

including phosphorylation, glycosylation, methylation, acetylation and a host of others, may be neglected. Some of this variation can be captured by mass spectrometric approaches (see review in this issue by Aebersold and Mann, page 198) and some by increasing the number of constructs that are generated for each gene. Another limitation to large-scale protein production is that the substantial class of membrane proteins is generally not amenable to the standardized procedures of genome-wide approaches.

Cloning of ORFs for subsequent expression requires a genomic set of gene-specific primers that is suitable for amplification by the polymerase chain reaction (PCR) and for subsequent insertion of the PCR products into appropriate plasmids. This latter requirement is met by use of forward and reverse primers that contain common 5' ends. The first example of this methodology was the genomic-scale PCR amplification of the ~6,000 *S. cerevisiae* ORFs for cloning into yeast plasmids⁴. A similar strategy was applied to more than 1,200 *Caenorhabditis elegans* ORFs predicted solely by a gene analysis programme, and ~70% of these were verified by sequence analysis of the PCR products⁵. Efforts are also under way for sets of mouse and human ORFs. With modern methods of high-throughput synthesis, primers can be made with high fidelity, at reasonable cost, and in 96-well format amenable for robotic manipulation. Insertion of the amplified ORFs into vectors generally uses any of several recombination-based methods that are now in widespread use (Box 1).

For biochemical analysis of proteins, their expression in a homologous system is ideal because the proteins are in their natural environment, are subject to native modifications, and can interact with their natural partners. This has been possible for proteins from yeast and from bacteria such as *Escherichia coli*, but heterologous expression is usually used for proteins from other organisms. In most cases, expression is attempted in *E. coli*, in which upwards of 60% of likely soluble proteins may be expressed in soluble form⁶. The most common alternative for expression is the use of insect cells, which results in modifications that are usually similar to mammalian cells. However, heterologous expression inevitably can lead to problems of expression and solubility for many proteins.

A primary goal of the genome-wide plasmid constructions is to incorporate a fusion tag, a short peptide or protein

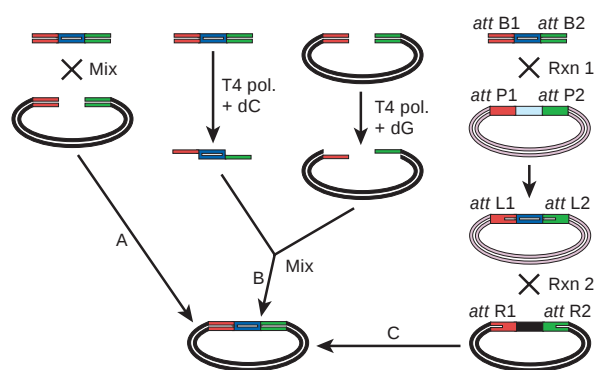
Box 1

Recombinational strategies for rapid and precise cloning of amplified DNA

For each strategy shown in the figure opposite, amplified open reading frame (ORF) DNAs (blue) all have a common 5' end sequence (red) and a different common 3' end sequence (green). The earliest method (strategy A) used gap repair-mediated recombination in yeast. Amplified ORFs are mixed with a linearized vector with ends that are identical to those of the amplified DNA, and the mixture is transformed directly into yeast, where gap repair-mediated recombination occurs *in vivo* with high efficiency. The result is precisely integrated DNA, which allows simple construction of in-frame fusions of genes.

Strategy B is ligation-independent cloning. Both 3' ends of amplified ORFs are constructed to lack a specific nucleotide (for example, dC), such that treatment with T4 DNA polymerase and dCTP removes the 3' ends until the first dC residue and leaves 5' overhangs of 12 or more bases. The corresponding plasmid with matching ends is treated with T4 DNA polymerase and dGTP, creating a corresponding overhang. After enzyme removal, DNAs are mixed and transformed directly into the bacterium *Escherichia coli*^{72,73}. This method has recently been used for the generation of 400 clones in a 3-day automated procedure⁷⁴. Thus, cloning the ORFs of a genome about the size of yeast might require only a few months of work.

Strategy C uses recombination by the Gateway cloning system of Invitrogen. This system is based on integration/excision of phage λ in *E. coli*. During integration, recombination between the λ att P site and the host att B site results in an integrated phage with ends called att L and att R. During excision, recombination between att L and att R sites regenerates att P and att B. In the Gateway application of this reaction, recombination is effected at each end of the DNA by pairs of att sites that are similar but not identical. Thus, amplified ORF DNAs with ends bearing att B1 and att B2 sequences recombine *in vitro* with a donor vector containing att P1 and att P2 sequences in the presence of components of λ integration to form the resulting entry plasmid, now bearing att L1 and att L2 sites. A second recombination



reaction *in vitro* with a destination vector containing att R1 and att R2 sites (black) and components of λ excision allows transfer of the ORF into the expression vector, regenerating att B sites. A significant advantage of this approach is that the cloned ORF DNA can be transferred easily from one plasmid to a number of other plasmids by a simple set of further recombination reactions *in vitro*. This approach has recently been applied for the analysis of ORFs from humans and *Caenorhabditis elegans*^{5,6}.

In yeast, a powerful alternative to plasmid expression of ORFs can be provided by chromosomal expression of tagged ORFs. This approach uses transformation of linear DNA into yeast with selection for recombination at a homologous chromosomal site. It has been used to construct a library of 1,548 strains⁴⁵ in which the coding sequence for a purification tag was inserted at the end of the ORFs. The resulting strains each express an ORF with a C-terminal fusion tag under control of its natural promoter, thereby ensuring as normal an expression pattern as possible, and stable propagation of the inserted gene.

domain that becomes linked to each member of a set of proteins. The use of these tags has continued to revolutionize biochemical analysis. For purification of biochemically active proteins, protein affinity tags (see Box 2) feature high affinity and selectivity for binding to specific resins to facilitate purification and elution under conditions that retain activity⁷. The recent application of genomic high-throughput purification illustrates the utility of such tags. Through the use of manual methods in 96-well format, 5,800 individual yeast glutathione S-transferase (GST) fusion proteins were purified 1,152 at a time and used successfully for biochemical analysis⁹. Current approaches now apply automation to parallel purification. But expression of each fusion protein and the purification of the corresponding tagged proteins require the use of a generic scheme. Inevitably, there will be members of a protein set that cannot be expressed, solubilized or purified under these generic conditions, because of the loss of a cofactor, inappropriate buffers, or other incompatible conditions. Additionally, proteins may be functionally inactive as fusion proteins.

Probing protein activity on a proteomic scale

The ultimate value of genomic sets of strains expressing tagged proteins, or of the corresponding purified proteins, is their potential for parallel analysis of the proteome. In this way one can, in principle, identify all of the proteins with a particular function or property in a single systematic experiment.

Biochemical genomics and functional protein microarrays

Two very different methods have been used to probe genomic sets of proteins for biochemical activity. One method has been termed a

biochemical genomics approach, which uses parallel biochemical analysis of a proteome comprised of pools of purified proteins in order to identify proteins and the corresponding ORFs responsible for a biochemical activity⁹. As applied to *S. cerevisiae*, this approach involved the generation of a set of 6,144 yeast strains, each expressing a distinct *S. cerevisiae* ORF as a GST-ORF fusion protein, followed by purification of the fusion proteins in pools. A biochemical activity is mapped to a specific ORF by assaying the pools for an activity, and then deconvoluting positive pools by preparation and analysis of subpools of the proteins. This method has been used to rapidly identify a number of yeast genes whose products co-purify with activities, including two proteins implicated in the metabolism of an NAD derivative produced during transfer RNA splicing⁹, a cytochrome *c* methyltransferase⁹, a tRNA dihydrouridine synthase¹⁰, both members of a tRNA m⁷G methyltransferase complex¹¹, and a new DNA-binding protein implicated in the transcriptional regulation of the yeast *SUC2* gene¹².

Important features of this approach include its speed at assigning catalytic function to ORFs, its generality for virtually any type of catalytic activity, and its sensitivity. High sensitivity is obtained both because the fusion proteins are overexpressed, and because background proteins are removed during purification. The lack of background proteins allows activities to be assayed for hours without destruction of product, substrate or proteins, yielding a huge increase in sensitivity for catalytic activities¹³. Additionally, it allows the detection of complexes of more than one protein, which otherwise cannot be detected by overproduction of a single component¹¹. Because the average protein in these preparations is present at

concentrations of ~20 nM, this approach is also suitable for the detection of protein–ligand complexes, which, unlike enzymatic activities, do not benefit from prolonged incubation¹³. The requirements of this method for a functional amino-terminal ORF fusion, and for effective solubilization and purification of the GST–ORF fusion proteins in active form, are often satisfied. However, the library has some bias against larger proteins, and those that retard growth during propagation^{13,14}.

The second approach for analysing genomic sets of proteins is the use of functional protein microarrays, in which individually purified proteins are separately spotted on a surface such as a glass slide and then analysed for activity. This approach has huge potential for rapid high-throughput analysis of proteomes and other large collections of proteins, and promises to transform the field of biochemical analysis (Fig. 1).

A critical first step in generating these arrays has been the development of general methods for arraying a genomic set of proteins on a solid surface without denaturing the proteins, and at high enough density for detection of activity. Recently, arrays have used both glass slides and chips with modified surfaces engineered to carry pads, films, nanowells or microfluidic channels^{8,15–20}. Although such modified surface structures require sophisticated engineering, they reduce evaporation and denaturation during drying, increase protein-binding capacity, and prevent cross-contamination because of the physical boundaries separating each sample.

A comprehensive microarray screening of a class of proteins was described by Zhu *et al.*¹⁶, who analysed the substrate specificities of 119 yeast protein kinases using 17 different test substrates that were adhered to the surface of nanowell microarrays. The experiments of MacBeath and Schreiber¹⁵ further demonstrated the potential of functional protein microarrays. In this study, proteins were tethered covalently to chemically activated glass slides, and then shown to be active for different classes of activities. Thus, three well-studied protein–protein interactions could be detected with fluorescently labelled protein probes, three different substrate proteins were shown to be phosphorylated specifically by protein kinases known to act on them, and three types of protein–small-molecule interactions could be detected in the micromolar range using small molecules bound to fluorescently labelled beads, which allows greater sensitivity owing to avidity effects. Finally, it was shown that a single protein could be detected at high resolution on a single glass slide in the midst of 10,799 identical spots of another protein. Taken together, Zhu *et al.*¹⁶ and MacBeath and Schreiber¹⁵ showed the huge potential of protein microarrays for parallel biochemical analysis.

The first full-scale genomic protein microarray was demonstrated by Zhu *et al.*⁸. In this experiment, 5,800 (94%) of the predicted yeast ORFs were cloned, and greater than 80% of these produced detectable amounts of protein, after purification in a high-throughput protocol. The proteins were spotted onto nickel-coated glass slides and used for the analysis of two different binding activities. First, a biotinylated calmodulin probe detected 6 of the known calmodulin-binding proteins that were present in the purified collection, as well as 33 new, potentially interacting proteins. Second, biotinylated liposomes detected 150 proteins that bind different phosphoinositides⁸. These experiments opened a new field in which entire proteomes can be screened for binding and other biochemical assays.

This approach can be extended in several different ways. Binding can be studied in real-time by use of a surface plasmon resonance (SPR) biosensor surface with 64 individual immobilized sites in a single flow cell, which can be scaled to 400 assays per day²¹. Peptides can also be analysed using microarrays. Recently, a monolayer-coated gold chip was shown to be useful for immobilization of peptides for biochemical analysis using detection by a phosphorimager, SPR and fluorescence microscopy²². Synthesis of peptide microarrays may become more practical with the development of methods for *in situ* synthesis of high-density peptide microarrays,

Box 2

Commonly used affinity tags

Each affinity tag used for purification of fused proteins has its unique features. Glutathione S-transferase binds tightly to glutathione-agarose and is eluted with glutathione, allowing very high levels of purification in one step; however, it is known to dimerize and this may interfere with protein function. His6 binds immobilized metal-ion columns and is eluted with imidazole; this tag is very simple, but it results in more modest purification than some other tags. Calmodulin-binding peptide binds calmodulin-agarose columns and is eluted with EGTA⁷⁵; purification is efficient, but proteins with required divalent metal ions may be affected.

An extremely useful variation of these commonly used tags uses very-high-affinity purification tags coupled with protease cleavage sites to remove the purification tag and simultaneously elute the proteins. An example is the generation of fusion proteins in which the target protein is fused to a tobacco etch virus (TEV) protease site linker and a protein A domain. The protein A domain binds tightly to immunoglobulin- γ columns, following which the target protein is released using treatment with the highly specific TEV protease⁷⁵. This results in extremely efficient purification of proteins.

Additional fusion tags used for other purposes include maltose-binding protein, which increases the solubility of fused proteins⁷⁶; epitope tags such as haemagglutinin, myc and FLAG tags, for immunoprecipitation and detection; and green fluorescent protein and its derivatives, which have been used for localizing proteins in living cells and for detection of protein interactions through fluorescence resonance energy transfer. Often tags or domains are used in concert. Examples include the use of two tags to effect very high purification of complexes of proteins for mass spectrometry analysis⁴⁵, the use of nuclear localization tags in concert with DNA-binding domains or transcription-activation domains for two-hybrid experiments, the use of purification tags coupled with a His6 tag for arraying on a microscope slide⁸, and the use of purification tags coupled with tags for immunodetection.

using photolithography or light-directed synthesis²³. Carbohydrate and small-molecule microarrays have also shown great potential for characterizing protein–small-molecule binding activities^{24,25}.

Both the biochemical genomics approach and protein microarrays have advantages and disadvantages. Use of biochemical genomics for yeast requires only 64 assays to cover the genome, is flexible for many types of assays, and is particularly useful for enzymatic activities. But use of pools does not allow easy assessment of the quality of each individual protein in a pool, can cause interference by the 95 other proteins present in the mixture, is not well-suited to binding assays with fluorescent probes, and cannot easily handle multiple positives at once. Use of microarrays to probe activity allows individual assessment of the quality of each protein, the immediate identification of the source ORF responsible for a particular activity, the identification of multiple positives in a single round, and high-throughput analysis of activities via automated arraying, assaying and scanning. However, it requires individual growth of 6,000 strains (for yeast) and 6,000 individual purifications of proteins, and is best suited at present for binding assays using fluorescent probes or activity assays with tethered substrates.

A second type of protein microarray, which is early in development, is the analytical microarray. Here, a genomic set of protein-specific ligands such as antibodies, nucleic acid aptamers or chemical probes is spotted on a microarray, and then the levels of different proteins in an extract are quantified in parallel by binding extract proteins to the microarray. Analytical protein microarrays are starting to realize their potential for monitoring protein expression on a proteome-wide scale and in medical diagnostics. Microarrays

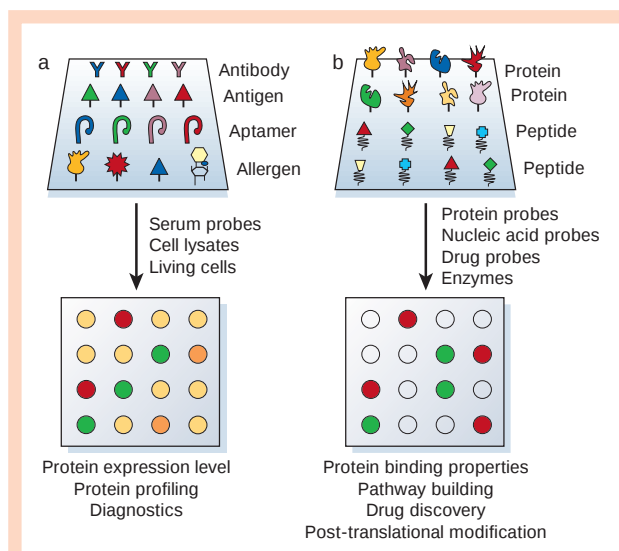


Figure 1 Analytical versus functional protein microarrays. **a**, Analytical protein microarray. Different types of ligands, including antibodies, antigens, DNA or RNA aptamers, carbohydrates or small molecules, with high affinity and specificity, are spotted down onto a derivatized surface. These chips can be used for monitoring protein expression level, protein profiling and clinical diagnostics. Similar to the procedure in DNA microarray experiments, protein samples from two biological states to be compared are separately labelled with red or green fluorescent dyes, mixed, and incubated with the chips. Spots in red or green colour identify an excess of proteins from one state over the other. **b**, Functional protein microarray. Native proteins or peptides are individually purified or synthesized using high-throughput approaches and arrayed onto a suitable surface to form the functional protein microarrays. These chips are used to analyse protein activities, binding properties and post-translational modifications. With the proper detection method, functional protein microarrays can be used to identify the substrates of enzymes of interest. Consequently, this class of chips is particularly useful in drug and drug-target identification and in building biological networks.

containing antibodies, antigens or in some cases peptides and other biomolecules have been used to monitor differential expression of proteins in colon carcinoma cells²⁶, cell-surface antigens specific for particular cell types²⁷, and autoantibodies in patient sera^{28,29}. The main problems with antibody-mediated analytical protein microarrays are specificity and quantitation. Most antibodies cross-react with proteins other than the antigen of interest, which leads to poor quantification. Haab and colleagues¹⁷ showed that only 23% of 115 well-characterized antibody–antigen pairs could be accurately quantified at the level of $1 \mu\text{g ml}^{-1}$ soluble antigen, although 60% of the binding interactions could be estimated qualitatively. Nonetheless, it seems likely that better and more efficient methods will be developed in the coming years to quantitatively assay the amounts of proteins in a high-throughput, parallel manner.

Other large-scale activity-based assays

Other activity assays have been used that address functional classes of activities within the proteome. The goal of one approach was to assess all of the DNA targets of the known DNA-binding protein regulators of yeast under one defined growth condition³⁰. To this end, a series of strains was constructed in which each of the 141 known yeast regulators was epitope-tagged at its carboxy terminus and expressed under control of its normal promoter at its appropriate chromosomal locus (see Box 1). After growth of each strain, chromatin immunoprecipitation analysis was carried out, in which each tagged protein was purified along with its population of bound DNA, and the identity and amount of the DNA was determined with conventional DNA microarrays. The technique was used with 106 of the 141 known

transcription factors, and the study allowed not only a genomic view of the regulatory modules of each gene, but also a description of a number of different networks of transcription regulation in the cell, and a functional assessment of the role of each transcription factor in yeast.

Another general method for assessing catalytic activity of the proteome is activity-based protein profiling. In this method, an extract is treated with a chemical probe that reacts covalently with any protein having a specific class of activity, and modified proteins are detected with a second tag such as biotin that is present on the reactive chemical^{31–33}. The key to the approach is the use of a probe that is specific for the activity, but general for all proteins with that class of activity. The method has been applied to probe cysteine proteases, resulting in the identification of two previously known caspase species in cells induced for apoptosis and evidence for several candidates in another cell line^{31,34}. It has also resulted in the identification of three previously known cathepsins and several other reactive proteins in rat kidney extracts, and demonstrated distinct labelling patterns during the progression of skin cancer in mice³⁵. Activity-based protein profiling has also been applied to probe serine hydrolases, resulting in the identification of two such hydrolases from rat brain and the detection of a number of others in different tissues³⁶. It is evident from these studies that this method is remarkably useful for profiling extracts to define the number of different activities of a particular type, the amounts of each protein in the active state, and the onset of the activity in different cell states.

Recently, this technology has been extended in three ways. First, a general isolation procedure was developed to purify and identify multiple reacted proteins in parallel. Denatured proteins were captured with avidin beads and then subjected to SDS-polyacrylamide gel electrophoresis, trypsin treatment and mass spectrometry³⁷. Second, a panel of different fluorescent derivatives of activity-based probes of the papain family of cysteine proteases was used to monitor active proteases in living cells, and to enable facile *in vivo* screening of small-molecule inhibitors for their activity and specificity³⁸. Third, small-molecule probes have been developed that are active against multiple types of enzymes, which allows profiling of several species simultaneously³⁹. The unique ability of activity-based protein profiling to monitor active species of a panel of enzymes in cells gives this method huge potential in profiling signal transduction pathways in development and differentiation, as demonstrated by the recent analysis of the activity, subcellular distribution and glycosylation state of the serine hydrolase superfamily in cancer cells⁴⁰.

A related activity-based probe involves the specific targeting of a single protein kinase *in vitro* or *in vivo* to elucidate its function. Identification of the natural targets of a protein kinase is of enormous importance because there are so many protein kinases in the proteome, a large fraction of the proteins in the cell are phosphorylated, and phosphorylation often has significant effects on protein function. To accomplish this, Shokat and colleagues⁴¹ re-engineered a highly conserved region of the ATP-binding site of protein kinases to allow the use of ATP analogues and kinase inhibitors that would not normally be active. Thus, a specific kinase can be retailed such that it alone is inhibited *in vivo*, allowing an assessment of its function⁴¹, or such that it is the only active kinase in extracts, allowing facile identification of potential substrates⁴². For example, the specifically activated kinase JNK was used to identify a new substrate in crude extracts by isolation of the corresponding phosphorylated protein from two-dimensional gels, followed by mass spectrometry⁴². This approach is generally applicable to many serine/threonine protein kinases and tyrosine protein kinases⁴³, and promises to have a prominent role in deducing the range and scope of function of this broad class of cellular activity.

Protein interaction analysis

One powerful method for deducing protein function is to identify the interacting partners of proteins, as proteins that interact with one

another or are part of the same complex are generally involved in the same cellular processes. As such, there have been intensive efforts in the past few years to identify protein–protein interaction on a large scale. Two types of approaches have been used: the two-hybrid system described below, which is used to detect binary interactions *in vivo*, and biochemical co-purification of complexes using affinity tags, coupled with protein identification using mass spectrometry, which defines the total spectrum of complexes for a particular tagged protein^{44,45}. The latter is reviewed by Aebersold and Mann on page 198 of this issue and will not be discussed. Fluorescent-based interaction assays have also been developed, but have not been used on a high-throughput basis.

Genome-wide two-hybrid approaches

The yeast two-hybrid assay⁴⁶ provides a genetic approach to the identification and analysis of protein–protein interactions. It relies on the modular nature of many eukaryotic transcription factors, which contain both a site-specific DNA-binding domain and a transcriptional-activation domain that recruits the transcriptional machinery. In this assay, hybrid proteins are generated that fuse a protein X to the DNA-binding domain and protein Y to the activation domain of a transcription factor (Fig. 2a). Interaction between X and Y reconstitutes the activity of the transcription factor and leads to expression of reporter genes with recognition sites for the DNA-binding domain. In the typical practice of this method, a protein of interest fused to the DNA-binding domain (the so-called ‘bait’) is screened against a library of activation-domain hybrids (‘preys’) to select interacting partners.

Key advantages of the two-hybrid assay are its sensitivity and flexibility. The sensitivity derives in part from overproduction of proteins *in vivo*, their designed direction to the nuclear compartment where the interactions are monitored, the large number of variable inserts of the interacting proteins that can be examined at once, and the potency of the genetic selections. This sensitivity leads to the detection of interactions with dissociation constants around 10^{-7} M, in the range of most weak protein interactions found in the cell, and is more sensitive than co-purification, which requires stability of a complex through dilution from cell lysis, and through subsequent purification steps. This sensitivity also allows detection of certain transient interactions or those that might affect only a subpopulation of the hybrid proteins.

Flexibility of the assay is provided by calibration to detect interactions of varying affinity by altering the expression levels of the hybrid proteins, the number and nature of the DNA-binding sites, and the composition of the selection media. Disadvantages of the yeast assay include the unavoidable occurrence of false negatives and false positives. False negatives include proteins such as membrane proteins and secretory proteins that are not usually amenable to a nuclear-based detection system, proteins that activate transcription when fused to a DNA-binding domain, proteins that fail to fold correctly, and interactions dependent on domains occluded in the fusions or on post-translational modifications. False positives include colonies not resulting from a bona fide protein interaction, as well as colonies resulting from a protein interaction not indicative of an association that occurs *in vivo*. Predominantly, false positives seem to be due to spurious transcription that does not derive from any interaction occurring between the hybrid proteins.

The two-hybrid system evolved to a proteomics strategy by the construction of ordered arrays of strains expressing either DNA-binding domain or activation-domain fusion proteins, the implementation of improved selection methods and plasmids, the use of mating to introduce pairs of plasmids for testing, and the use of automation.

Different genome-wide two-hybrid strategies have been used to analyse protein interactions in *S. cerevisiae*. One approach involved screening a large number of individual proteins against a comprehensive library of randomly generated fragments (Fig. 2b), as was

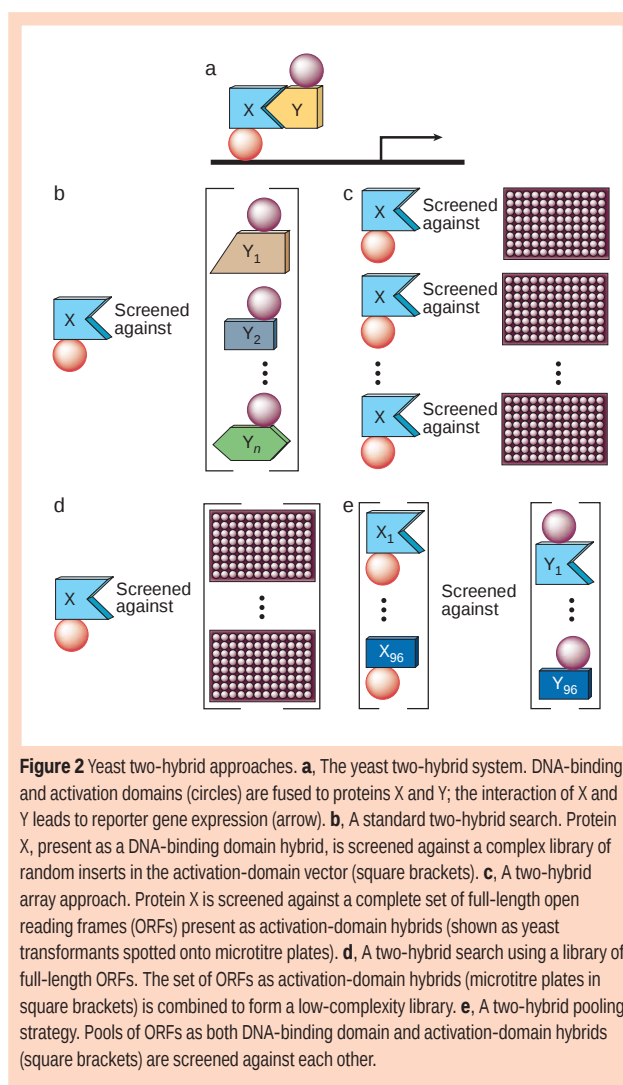


Figure 2 Yeast two-hybrid approaches. **a**, The yeast two-hybrid system. DNA-binding and activation domains (circles) are fused to proteins X and Y; the interaction of X and Y leads to reporter gene expression (arrow). **b**, A standard two-hybrid search. Protein X, present as a DNA-binding domain hybrid, is screened against a complex library of random inserts in the activation-domain vector (square brackets). **c**, A two-hybrid array approach. Protein X is screened against a complete set of full-length open reading frames (ORFs) present as activation-domain hybrids (shown as yeast transformants spotted onto microtitre plates). **d**, A two-hybrid search using a library of full-length ORFs. The set of ORFs as activation-domain hybrids (microtitre plates in square brackets) is combined to form a low-complexity library. **e**, A two-hybrid pooling strategy. Pools of ORFs as both DNA-binding domain and activation-domain hybrids (square brackets) are screened against each other.

used to identify numerous interactions for proteins implicated in RNA splicing⁴⁷. A second approach used systematic one-by-one testing of every possible combination of proteins using a mating assay with a comprehensive array of strains. In this way, 192 baits were screened against an array of essentially all activation-domain fusions of full-length yeast ORFs to identify 281 putative interactions⁴⁸, and ~1,000 proteins have been screened to date (S.F., unpublished data). A third approach used a one-by-many mating strategy in which each member of a nearly complete set of strains expressing yeast ORFs as DNA-binding domain hybrids was mated to a library of strains containing activation-domain fusions of full-length yeast ORFs (Fig. 2d), resulting in 692 positives⁴⁸. A fourth variation involved mating of defined pools of strain arrays⁴⁹. This approach required cloning all of the yeast ORFs into both two-hybrid vectors, followed by pooling sets of 96 transformants each. Matings were conducted for the 62×62 combinations of pools, and positives were sequenced (Fig. 2e), resulting in a total of 4,549 positives, of which the 841 that were identified more than three times form a core data set.

In addition to the analyses of yeast proteins, large-scale two-hybrid studies have been carried out for proteins of *Helicobacter pylori*⁵⁰, *C. elegans*⁵¹ and *Drosophila melanogaster* (R. Finley, personal communication).

Notably, these approaches are not exclusive; for example, full-length ORFs are often used in screens of random libraries, and protein fragments can be tested in a one-by-one format against an

activation-domain array. Compared to systematic mating, random insert or defined ORF libraries require more statistical sampling to ensure adequate coverage of the interactions. They also require sequencing of plasmids to identify interacting partners and tend, on average, to yield fewer interactions than systematic mating, although throughput is faster. Random fragment libraries may also reveal domains that might be masked, and smaller fusion proteins work better in the assay and provide direct information about interaction domains.

Unlike the case for a single two-hybrid experiment conducted by an individual laboratory dedicated to the investigation of a specific biological question, the proteomic two-hybrid projects produce potential interactions at a rate too rapid to allow individual testing for confirmation. Small-scale experiments generally allow the elimination of false positives, yielding a literature focused on a few interactions that have often been validated by additional experimentation; by contrast, genome-wide projects necessarily report all of their putative interactions. This raises the question of the accuracy of genomic data in general, and of two-hybrid data in particular.

Several analyses of genomic two-hybrid results suggest that about 50% are correct^{52–56}. These studies have set the tone for how other large proteomic data sets can be mined to retrieve biologically significant findings. For example, one approach is based on the fact that genes encoding proteins involved in the same function tend to be co-expressed⁵³. A second strategy⁵³ assesses reliability by determining whether two proteins that interact putatively have paralogues that also interact. A third uses information about protein localization (that is, which proteins lie in the same subcellular compartment) to increase the accuracy of the two-hybrid interaction data⁵⁴. These analyses indicate that the data from small-scale studies are of considerably greater reliability than that from high-throughput studies. Additionally, they show how computational assessment of large-scale data that relies on a different property of proteins can find the most reliable interactions (for example, Deane *et al.*⁵³ identified ~1,400 interactions of yeast proteins that are likely to be correct). Computational analysis also indicates that experimental corroboration of protein interactions by a combination of methods is likely to yield data that are substantially more reliable^{54,56,57}. Finally, the large number of false negatives in proteomic studies suggests that most of the studies completed so far are far from saturated and that the universe of protein–protein interactions is likely to be several times higher than those currently known.

The principle of using hybrid proteins to analyse interactions has been extended to examine DNA–protein interactions, RNA–protein interactions, small-molecule–protein interactions, and interactions dependent on bridging proteins or post-translational modifications⁵⁸. Additionally, the reconstitution of proteins other than transcription factors, such as ubiquitin, has been used to establish reporter systems to detect interactions⁵⁸, and these may enable the analysis of proteins not generally suitable for the traditional two-hybrid assay, such as membrane proteins. Although some of these alternative methods may be robust enough for high-throughput proteomic analysis, so far most of these approaches have been demonstrated only for their initial proof of principle, or in screens of a small number of proteins.

Analysing protein interactions by fluorescence methods

Another potentially general method to detect protein–protein interactions involves the use of fluorescence resonance energy transfer (FRET) between fluorescent tags on interacting proteins. FRET is a non-radiative process whereby energy from an excited donor fluorophore is transferred to an acceptor fluorophore that is within ~60 Å of the excited fluorophore⁵⁹. After excitation of the first fluorophore, FRET is detected either by emission from the second fluorophore using appropriate filters, or by alteration of the fluorescence lifetime of the donor. Two fluorophores that are commonly used are variants of green fluorescent protein (GFP): cyan fluorescent protein (CFP)

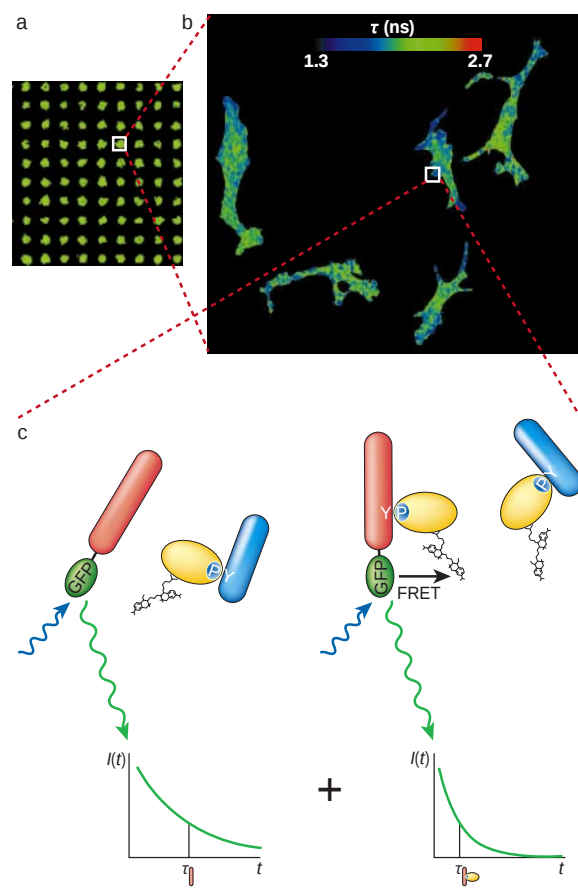


Figure 3 Principle of optical detection of protein post-translational modifications on a cell microarray. **a**, Local transfection techniques enable the expression of defined green fluorescent protein (GFP)–cDNA in clusters of cells at each location. **b**, Fluorescence lifetime imaging microscopy on the cell clusters reveals fluorescence resonance energy transfer (FRET) between GFP and an acceptor dye on a protein module or antibody directed against phosphorylated tyrosines by a drop in the fluorescence lifetime (τ) of the GFP (blue areas). **c**, Binding of the acceptor-tagged protein module or antibody (yellow) to the phosphorylated amino acid (YP) is detected only on the GFP fusion protein (red) via FRET and not on other phosphorylated proteins (blue). The extent of phosphorylation on the GFP fusion protein at each optically resolvable volume element in the cell can be determined from the linear combination of the GFP fluorescence decays in the presence and absence of FRET.

and yellow fluorescent protein (YFP)⁶⁰. A number of protein interactions have been demonstrated in cells by FRET microscopy⁵⁹, including oligomerization of the Fas receptor⁶¹, interaction between the apoptosis-regulating proteins Bcl-2 and Bax in mitochondria⁶², and interaction between Pit-1 and Ets-1 transcription factors in the nucleus⁶³.

The potential of FRET is considerable, for two reasons. First, it can be used to make measurements in living cells, which allows the detection of protein interactions at the location in the cell where they normally occur, in the presence of the normal cellular milieu. For example, inducible interactions have been demonstrated, such as the binding of Grb2 to activated epidermal growth factor receptors⁶⁴ and the hormone-induced binding of co-activator proteins to nuclear receptors⁶⁵. Second, transient interactions can be followed with high temporal resolution in single cells.

In principle, one can imagine two classes of high-throughput FRET screens that might be used. First, protein interactions within the proteome might be mapped by performing FRET screens on cell

arrays that are co-transfected with complementary DNAs bearing CFP and YFP fusion proteins. In practice, however, this may be difficult because of the high incidence of false negatives. These can arise from the lack of proper geometric orientation for FRET detection, and from the low FRET contributions in the fluorescence signals, which are difficult to detect above the background fluorescence from direct acceptor excitation or donor emission, particularly when expression levels of donor and acceptor tagged proteins are unbalanced.

Second, post-translational modifications might be detected by challenging GFP-cDNA donors with a FRET acceptor-tagged protein specific for that class of modification⁵⁹. For example, cell microarrays expressing GFP-cDNA fusion libraries can be permeabilized and incubated with an anti-phosphotyrosine antibody conjugated to a FRET acceptor to measure tyrosine phosphorylation of any of the GFP fusion proteins by fluorescence lifetime imaging of the donor⁵⁹. This approach allows specific detection of the signal even though the antibody binds all phosphotyrosine-containing proteins (Fig. 3). And because the acceptor fluorescence is filtered out in this approach, it permits the use of saturating amounts of labelled acceptor molecules. To boost the signal from such an experiment, the FRET acceptor-tagged protein can be tagged with several acceptor fluorophores. If this method becomes practical, similar approaches could be used to monitor other post-translational modifications.

Protein localization

A proteomics strategy of increasing importance involves the localization of proteins in cells as a necessary first step towards understanding protein function in complex cellular networks. A proteome-scale analysis of protein localization has been performed in *S. cerevisiae* by immunolocalization of epitope-tagged gene products⁶⁶. These experiments established the subcellular localization of 2,744 proteins, 955 of which had no previously known function. The data were integrated with those previously published to identify the localization of 55% of the yeast proteome, which was extended to the full proteome by using a Bayesian estimation system⁶⁶. This study corroborated that there is a good correlation between protein function and localization in the cell.

The discovery of GFP and the development of its spectral variants⁶⁰ has opened the door to analysis of proteins in living cells by use of the light microscope. Large-scale approaches of localizing GFP-tagged proteins in cells have been performed in the genetically amenable yeast *S. pombe*^{67,68} and in *Drosophila*⁶⁹. For the localization of proteins in mammalian cells, a strategy was developed that enables the systematic GFP tagging of ORFs from novel full-length cDNAs that are identified in genome projects⁷⁰. This approach proved remarkably successful, showing a high correlation between prediction and the subsequent subcellular localization of targeted proteins, and could be fully automated.

The parallel functional analysis of many proteins in cells has become possible by a microarray-driven gene expression system⁷¹. In this system, mammalian cells are cultured on glass slides printed in defined locations with different DNAs specifying, for example, different defined cDNA-GFP fusions. The local transfections of cells growing over the DNA spots allow the simultaneous observation of many different fusion constructs, which can be correlated with the coordinates to link the images with the identity of any particular DNA. In principle, this approach can be applied to steady-state imaging for localization, to dynamic imaging to monitor changes during signal transduction, and to FRET to monitor changes in interactions.

Outlook

The promise of proteomics is the precise definition of the function of every protein in the cell, and how that function changes in different environmental conditions, with different modification states of the protein, in different cellular locales, and with different interacting partners. Just in the past few years, tremendous progress has been

made in dissecting the functions of proteins using a battery of newly developed, sophisticated genome-wide approaches. Yet there is still a need both for additional high-throughput technologies and for computational methods to analyse large data sets and to integrate complex and disparate kinds of protein information. Another challenge will be for the proteomics community to work hand in hand with those focused on biological problems in order to best convert the broad but shallow proteomic data into deeper understanding. Within the next decade, we might have a reasonably complete picture of the proteome of a simple model organism such as yeast. This picture, in turn, will provide a blueprint for understanding the proteomes of other more complex model organisms and of humans. □

doi:10.1038/nature01512

- Kenyon, G. L. *et al.* Defining the mandate of proteomics in the post-genomics era. Workshop Report: National Academy of Sciences, Washington DC, USA. *Mol. Cell. Proteomics* **1**, 763–780 (2002).
- Cliften, P. F. *et al.* Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11**, 1175–1186 (2001).
- Kumar, A. *et al.* An integrated approach for finding overlooked genes in yeast. *Nature Biotechnol.* **20**, 58–63 (2002).
- Hudson, J. R. Jr *et al.* The complete set of predicted genes from *Saccharomyces cerevisiae* in a readily usable form. *Genome Res.* **7**, 1169–1173 (1997).
- Reboul, J. *et al.* Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nature Genet.* **27**, 332–336 (2001).
- Braun, P. *et al.* Proteome-scale purification of human proteins from bacteria. *Proc. Natl Acad. Sci. USA* **99**, 2654–2659 (2002).
- Nilsson, J., Stahl, S., Lundberg, J., Uhlen, M. & Nygren, P. A. Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. *Protein Exp. Purif.* **11**, 1–16 (1997).
- Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
- Martzen, M. R. *et al.* A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286**, 1153–1155 (1999).
- Xing, F., Martzen, M. R. & Phizicky, E. M. A conserved family of *Saccharomyces cerevisiae* synthases effects dihydrouridine modification of tRNA. *RNA* **8**, 370–381 (2002).
- Alexandrov, A. V., Martzen, M. R. & Phizicky, E. M. Two proteins that form a complex are required for 7-methylguanosine modification of yeast tRNA. *RNA* **8**, 1253–1266 (2002).
- Hazbun, T. R. & Fields, S. A genome-wide screen for site-specific DNA-binding proteins. *Mol. Cell. Proteomics* **1**, 538–543 (2002).
- Phizicky, E. M. *et al.* Biochemical genomics approach to map activities to genes. *Methods Enzymol.* **350**, 546–559 (2002).
- Grayhack, E. J. & Phizicky, E. M. Genomic analysis of biochemical function. *Curr. Opin. Chem. Biol.* **5**, 34–39 (2001).
- MacBeath, G. & Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760–1763 (2000).
- Zhu, H. *et al.* Analysis of yeast protein kinases using protein chips. *Nature Genet.* **26**, 283–289 (2000).
- Haab, B. B., Dunham, M. J. & Brown, P. O. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* **2**, RESEARCH0004.1–0004.13 (2001).
- Zhu, H. & Snyder, M. Protein arrays and microarrays. *Curr. Opin. Chem. Biol.* **5**, 40–45 (2001).
- Weng, S. *et al.* Generating addressable protein microarrays with PROfusion covalent mRNA-protein fusion technology. *Proteomics* **2**, 48–57 (2002).
- Templin, M. F. *et al.* Protein microarray technology. *Trends Biotechnol.* **20**, 160–166 (2002).
- Myszka, D. G. & Rich, R. L. Implementing surface plasmon resonance biosensors in drug discovery. *Pharmacol. Sci. Technol. Today* **3**, 310–317 (2000).
- Houseman, B. T., Huh, J. H., Kron, S. J. & Mrksich, M. Peptide chips for the quantitative evaluation of protein kinase activity. *Nature Biotechnol.* **20**, 270–274 (2002).
- LeProust, E. *et al.* Digital light-directed synthesis. A microarray platform that permits rapid reaction optimization on a combinatorial basis. *J. Comb. Chem.* **2**, 349–354 (2000).
- Wang, D., Liu, S., Trummer, B. J., Deng, C. & Wang, A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nature Biotechnol.* **20**, 275–281 (2002).
- Kuruvilla, F. G., Shamji, A. F., Sternson, S. M., Hergenrother, P. J. & Schreiber, S. L. Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature* **416**, 653–657 (2002).
- Sreekumar, A. *et al.* Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. *Cancer Res.* **61**, 7585–7593 (2001).
- Belov, L., de la Vega, O., dos Remedios, C. G., Mulligan, S. P. & Christopherson, R. I. Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. *Cancer Res.* **61**, 4483–4489 (2001).
- Joos, T. O. *et al.* A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics. *Electrophoresis* **21**, 2641–2650 (2000).
- Robinson, W. H. *et al.* Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature Med.* **8**, 295–301 (2002).
- Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
- Faleiro, L., Kobayashi, R., Fearnhead, H. & Lazebnik, Y. Multiple species of CPP32 and Mch2 are the major active caspases present in apoptotic cells. *EMBO J.* **16**, 2271–2281 (1997).
- Cravatt, B. F. & Sorenson, E. J. Chemical strategies for the global analysis of protein function. *Curr. Opin. Chem. Biol.* **4**, 663–668 (2000).
- Adam, G. C., Sorenson, E. J. & Cravatt, B. F. Chemical strategies for functional proteomics. *Mol. Cell. Proteomics* **1**, 781–790 (2002).

34. Martins, L. M. *et al.* Activation of multiple interleukin-1 β converting enzyme homologues in cytosol and nuclei of HL-60 cells during etoposide-induced apoptosis. *J. Biol. Chem.* **272**, 7421–7430 (1997).
35. Greenbaum, D., Medzihradszky, K. F., Burlingame, A. & Bogoy, M. Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chem. Biol.* **7**, 569–581 (2000).
36. Liu, Y., Patricelli, M. P. & Cravatt, B. F. Activity-based protein profiling: the serine hydrolases. *Proc. Natl Acad. Sci. USA* **96**, 14694–14699 (1999).
37. Kidd, D., Liu, Y. & Cravatt, B. F. Profiling serine hydrolase activities in complex proteomes. *Biochemistry* **40**, 4005–4015 (2001).
38. Greenbaum, D. *et al.* Chemical approaches for functionally probing the proteome. *Mol. Cell. Proteomics* **1**, 60–68 (2002).
39. Adam, G. C., Sorensen, E. J. & Cravatt, B. F. Proteomic profiling of mechanistically distinct enzyme classes using a common chemotype. *Nature Biotech.* **20**, 805–809 (2002).
40. Jessani, N., Liu, Y., Humphrey, M. & Cravatt, B. F. Enzyme activity profiles of the secreted and membrane proteome that depict cancer cell invasiveness. *Proc. Natl Acad. Sci. USA* **99**, 10335–10340 (2002).
41. Bishop, A. C. *et al.* A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature* **407**, 395–401 (2000).
42. Habelhah, H. *et al.* Identification of new JNK substrate using ATP pocket mutant JNK and a corresponding ATP analogue. *J. Biol. Chem.* **276**, 18090–18095 (2001).
43. Bishop, A. C., Buzko, O. & Shokat, K. M. Magic bullets for protein kinases. *Trends Cell Biol.* **11**, 167–172 (2001).
44. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
45. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
46. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246 (1989).
47. Fromont-Racine, M. *et al.* Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **17**, 95–110 (2000).
48. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
49. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574 (2001).
50. Rain, J. C. *et al.* The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).
51. Walhout, A. J. M. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
52. Mrowka, R., Patzak, A. & Herzel, H. Is there a bias in proteome research? *Genome Res.* **11**, 1971–1973 (2001).
53. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1**, 349–356 (2002).
54. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein–protein interactions. *Genome Res.* **12**, 37–46 (2002).
55. Kemmeren, P. *et al.* Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* **9**, 1133–1143 (2002).
56. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
57. Edwards, A. *et al.* Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **18**, 529–536 (2002).
58. Fashena, S. J., Serebriiskii, I. & Golemis, E. A. The continued evolution of two-hybrid screening approaches in yeast: how to outwit different preys with different baits. *Gene* **250**, 1–14 (2000).
59. Wouters, F. S., Verveer, P. J. & Bastiaens, P. I. H. Imaging biochemistry inside cells. *Trends Cell Biol.* **11**, 203–211 (2001).
60. Tsien, R. Y. The green fluorescent protein. *Annu. Rev. Biochem.* **67**, 509–544 (1998).
61. Siegel, R. M. *et al.* Fas preassociation required for apoptosis signaling and dominant inhibition by pathogenic mutations. *Science* **288**, 2354–2357 (2000).
62. Mahajan, N. *et al.* Bcl-2 and Bax interactions in mitochondria probed with green fluorescent protein and fluorescence resonance energy transfer. *Nature Biotechnol.* **16**, 547–552 (1998).
63. Day, R. N. Visualization of Pit-1 transcription factor interactions in the living cell nucleus by fluorescence resonance energy transfer microscopy. *Mol. Endocrinol.* **12**, 1410–1419 (1998).
64. Sorkin, A., McClure, M., Huang, F. & Carter, R. Interaction of EGF receptor and Grb2 in living cells visualized by fluorescence resonance energy transfer (FRET) microscopy. *Curr. Biol.* **10**, 1395–1398 (2000).
65. Llopis, J. *et al.* Ligand-dependent interactions of coactivators steroid receptor coactivator-1 and peroxisome proliferator-activated receptor binding protein with nuclear hormone receptors can be imaged in live cells and are required for transcription. *Proc. Natl Acad. Sci. USA* **97**, 4363–4368 (2000).
66. Kumar, A. *et al.* Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719 (2002).
67. Ding, D. Q. *et al.* Large-scale screening of intracellular protein localization in living fission yeast cells by the use of a GFP-fusion genomic DNA library. *Genes Cells* **5**, 169–190 (2000).
68. Sawin, K. E. & Nurse, P. Identification of fission yeast nuclear markers using random polypeptide fusion with green fluorescent protein. *Proc. Natl Acad. Sci. USA* **94**, 15146–15151 (1996).
69. Morin, X., Daneman, R., Zavortink, M. & Chia, W. A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98**, 15050–15055 (2001).
70. Simpson, J. C., Wellenreuther, R., Poustka, A., Pepperkok, R. & Wiemann, S. Systematic subcellular localisation of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* **1**, 287–292 (2000).
71. Ziauddin, J. & Sabatini, D. M. Microarrays of cells expressing defined cDNAs. *Nature* **411**, 107–110 (2001).
72. Aslanidis, C. & de Jong, P. J. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* **18**, 6069–6074 (1990).
73. Aslanidis, C., de Jong, P. J. & Schmitz, G. Minimal length requirement of the single-stranded tails for ligation-independent cloning (LIC) of PCR products. *PCR Methods Appl.* **4**, 172–177 (1994).
74. Dieckman, L., Gu, M., Stols, L., Donnelly, M. I. & Collart, F. R. High throughput methods for gene cloning and expression. *Protein Exp. Purif.* **25**, 1–7 (2002).
75. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17**, 1030–1032 (1999).
76. Kapust, R. B. & Waugh, D. S. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8**, 1668–1674 (1999).

Acknowledgements We thank T. Davis and E. Grayhack for comments on the manuscript. This work was supported by grants from the National Center for Research Resources and National Human Genome Research Institute of the National Institutes of Health. S.F. is an investigator of the Howard Hughes Medical Institute.

From words to literature in structural proteomics

Andrej Sali*, Robert Glaeser†, Thomas Earnest‡ & Wolfgang Baumeister§

*Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California, San Francisco, California 94143, USA

†Department of Molecular and Cell Biology, Stanley/Donner ASU, University of California, Berkeley, California 94720, USA

‡Berkeley Center for Structural Biology, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

§Department of Structural Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18 a, 82152 Martinsried, Germany (e-mail: baumeist@biochem.mpg.de)

Technical advances on several frontiers have expanded the applicability of existing methods in structural biology and helped close the resolution gaps between them. As a result, we are now poised to integrate structural information gathered at multiple levels of the biological hierarchy — from atoms to cells — into a common framework. The goal is a comprehensive description of the multitude of interactions between molecular entities, which in turn is a prerequisite for the discovery of general structural principles that underlie all cellular processes.

The structures of individual macromolecules are often uninformative about function if taken out of context. Just as words must be assembled into sentences, paragraphs, chapters and books to make sense, vital cellular functions are performed by structured ensembles of proteins (that is, complexes), not by freely diffusing and occasionally colliding proteins¹. Frequently, these complexes comprise ten or more subunits (Fig. 1). Recent proteomics studies with yeast, for example, have indicated that the number of complexes that exist at least transiently in a cell has been underestimated. The techniques of isolation and purification that are traditionally used in biochemistry tend to select for the most robust complexes, whereas the more weakly interacting and transient complexes escape attention and, therefore, analysis.

In recent years, two trends have emerged in structural biology: efforts to achieve a comprehensive coverage of individual protein structures (so-called structural genomics) and efforts to analyse the structures of large complexes^{2,3}. Structural biology has flourished in the wake of technological innovations in fields as diverse as biochemistry, molecular biology, computational biology, computer hardware and software, nuclear magnetic resonance (NMR) magnets and optimized pulse sequences, and synchrotron radiation, as well as advances in light and electron microscopy (EM) instrumentation and in detector technology. Notwithstanding the value and importance of the individual techniques, a combination of approaches is likely to be more powerful than any single method alone. In this review we discuss some integrated strategies and tactics that can be used for characterizing molecular complexes and for describing their interactions in a cellular context.

The challenge of myriads of complexes

Given the average length of 466 residues for a yeast protein and 173 residues for a domain in the CATH database⁴ (a hierarchical classification of protein domain structures), one can estimate that, on average, a protein is folded into approximately two domains. In the evolution of proteins,

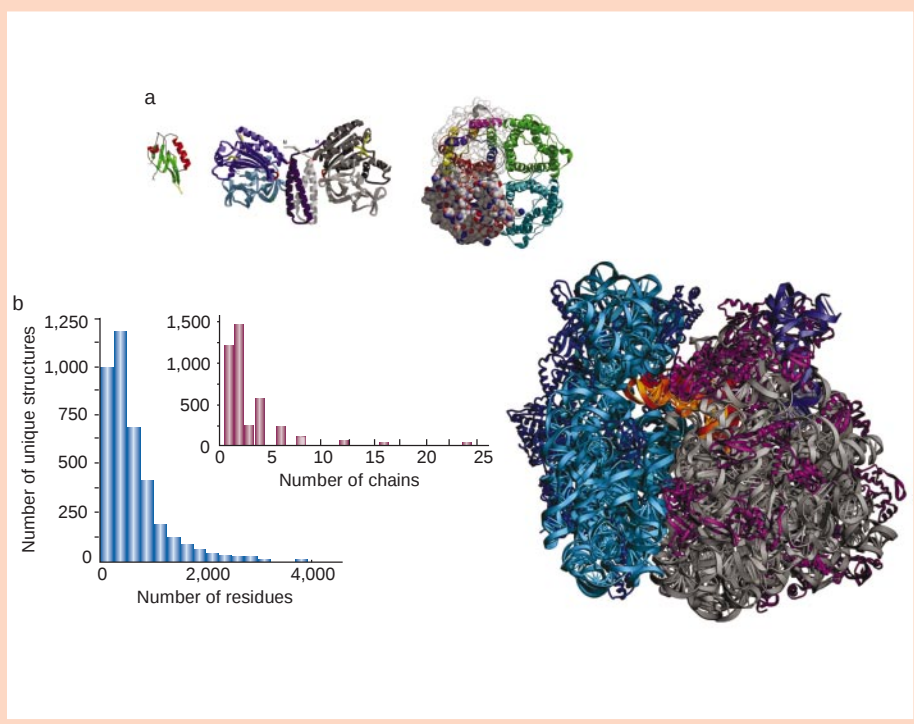
domains are important units that are shuffled, duplicated, and fused into larger proteins. Although the universe of distinct amino acid sequences is essentially unlimited, the number of different folding patterns for the domains is not. Extrapolation based on the existing databases of protein sequence and structure indicates that most of the natural domain sequences assume one of a few thousand folds⁵, of which ~1,000 are already known⁴.

In contrast to the folds, there are no satisfactory estimates of the number of different non-covalent macromolecular complexes with a unique structure and biological function. Such estimates are non-trivial to make because of the multitude of the component types (for example, proteins and nucleic acids), and the varying lifespan of the complexes (for example, transient complexes such as those involved in signalling, and stable complexes such as the ribosome). In addition, there is no self-evident definition of what is a 'complex' and whether two complexes are of different types. In an extreme view, a whole cell or even an organism may be seen as a single giant complex.

The Protein Quaternary Structure (PQS) database currently contains ~10,000 structurally defined protein assemblies of presumed biological significance, derived from a variety of organisms (<http://pqs.ebi.ac.uk/pqs-doc.shtml>); each assembly consists of at least two protein chains. Just like the folds, these assemblies can be organized into ~3,000 groups such that the members of the same assembly group share more than 30% sequence identity between the equivalent constituent protein chains (Fig. 1).

The most comprehensive information about both stable and transient protein complexes exists for the yeast proteome of ~6,200 proteins. But even for this model genome, uncertainties in the number, types and sizes of the complexes arise because of the difficulty in unravelling physical interactions from functional links⁶, binary from multiple physical interactions, transient from stable interactions, and direct interactions from indirect physical interactions through intermediates. In addition, each method may be impacted differently by the localization of the proteins in the cellular environment and may have significantly different rates of false positives and negatives.

Figure 1 Illustration of the size range of biomolecular structures solved by X-ray crystallography and the size distribution of structures contained in the Protein Quaternary Structure (PQS) database (<http://pqs.ebi.ac.uk>). **a**, X-ray crystallography can deal with a wide range of complexity. From top left to right, structures of: the PDZ domain of dishevelled, a molecular recognition domain that leads to protein–protein interactions; CheA, a dimeric multidomain bacterial signalling molecule; aquaporin, which serves as a transmembrane water channel; and 70S ribosome, which is the molecular machine for protein biosynthesis. **b**, The main histogram shows the distribution of the size of the entries in the PQS database. The 15,190 entries with at least one protein chain of at least 30 residues, when compared with each other, produced 3,876 clusters with more than 30% sequence identity and less than 30-residue length difference among the members within the same cluster. The inset shows the distribution of the number of chains in the representative structures for each group. As expected, the structures of large complexes are under-represented, given an estimated average size of a yeast complex of 7.5 proteins.



The Munich Information Center for Protein Sequences (MIPS)⁷ and Yeast Proteome Database (YPD)⁸ list ~11,000 binary interactions and functional links documented by focused, small-scale experiments⁹, corresponding on average to ~3.5 partners per protein. Large-scale yeast two-hybrid data^{10,11} indicate 1.7 partners per protein, when artefactual interactions are removed from consideration¹². On the other hand, the affinity purification of 1,739 yeast protein baits indicated 232 distinct complexes of an average size of 7.5 proteins, suggesting that the whole yeast proteome may contain ~900 complexes¹³. A comparison of these purified complexes against the complexes of known structure revealed that most of them are stable as opposed to transient, whereas the reverse applies to the interactions detected by the yeast two-hybrid methods^{14–17}. Only one-third of the binary interactions and functional links obtained by more than one high-throughput method occur in the curated MIPS/YPD set of the ~11,000 binary interactions and links, suggesting that the lower bound on the binary protein–protein interactions and functional links in yeast is ~30,000 (refs 9,18). This number corresponds to ~9 protein partners per protein or 3.6 protein partners per domain, not necessarily all direct or at the same time.

The human proteome may have an order of magnitude more complexes than the yeast cell; and the number of different complexes across all relevant genomes may be several times larger still. Therefore, there may be thousands of biologically relevant macromolecular complexes whose structures are yet to be characterized¹⁹.

Towards an unabridged dictionary of proteins

Currently, X-ray crystallography is the most prolific technique for the structural analysis of proteins and protein complexes, and it still is the ‘gold standard’ in terms of accuracy. While this technique has provided the majority of structures in the database of biomolecular structures, the fraction determined by NMR spectroscopy is also significant (currently 14%)²⁰. From the earliest structures of myoglobin and haemoglobin through the recent studies of RNA polymerase²¹, the ribosomal subunits^{22–24}, and the complete ribosome and its functional complexes²⁵, these structural data have contributed tremendously to our understanding of biology at the molecular level. As seen in Fig. 1, the sizes of the structures determined by X-ray crystallography range

from small proteins, such as the 100-residue PDZ domain, which recognizes and binds other proteins, to the 70S ribosome, which consists of 52 proteins and 3 RNA molecules, and has a relative molecular mass of ~2,500,000 (M_r 2,500K).

Crystallography requires that milligram quantities of a pure and monodisperse protein can be prepared, and that the protein can be

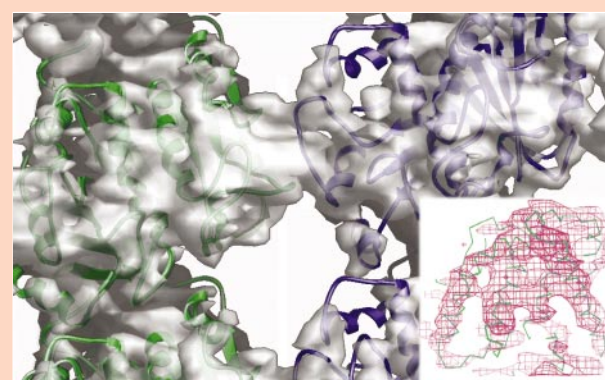


Figure 2 Docking the atomic model of tubulin into the cryo-EM density map of the assembled microtubule. The atomic model of tubulin, represented by its ribbon diagram, is shown docked into the 3D density of an intact, 13-protofilament microtubule, represented by the grey, transparent surface of the protein. The atomic model of 2D crystals of tubulin was refined at a resolution of 3.5 Å (ref. 38), and the model was docked as a rigid body into the microtubule density, which was obtained at a resolution of 8 Å by applying single-particle averaging methods to very short segments of ice-embedded microtubules⁴¹. The docked (hybrid) model shows which residues are responsible for forming the lateral contacts between individual protofilaments, information that could not be deduced from the structure of the protofilament alone. A short helix (upper centre) that is well ordered in the crystal structure is also shown to be disordered (lacking density) in the microtubule. The high precision with which this docking is specified by the data is shown more clearly in the insert, where the atomic model (represented by the C α backbone) is embedded within its corresponding portion of the 3D density (represented by the wire basket volume).

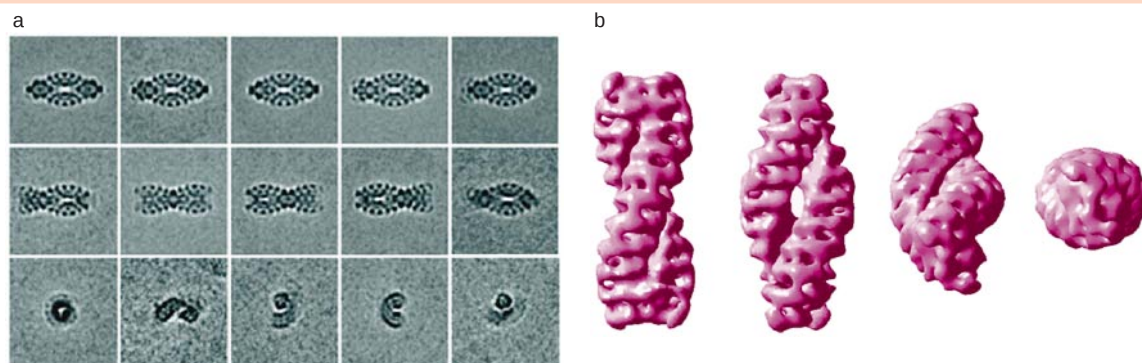


Figure 3 Representative example that illustrates the type of 3D reconstructions that can be obtained with large macromolecular complexes by single-particle cryo-EM. In this example, the specimen is a giant assembly of *Drosophila melanogaster* tripeptidyl peptidase II (TPP II). The protein monomers have a relative molecular mass of 150,000 (M_r 150K) and the intact assembly has a particle mass of ~6,600K (ref. 42). **a**, Some of the distinct views that are obtained by averaging many equivalent projections of individual particles randomly

oriented within a thin film of vitreous ice. A wide variety of side views can be distinguished, corresponding to rotation of the particle around its long axis. In addition, other projections, shown in the bottom row, correspond to particles that are viewed directly on axis or at a small tilt relative to the axial view. **b**, 3D surface representation of the TPP II complex at 3.3-nm resolution, in which the particle is first rotated about its long axis and then it is tilted to bring the long axis perpendicular to the page.

induced to form three-dimensional (3D) periodic arrays (that is, crystals). Therefore, almost all proteins used for structural studies are expressed in heterologous expression systems. Bacterial expression systems are simple and rapid, in addition to being amenable to incorporation of selenium as an anomalous scatterer for determining phases. However, overexpression in bacteria may not produce large amounts of the correctly folded protein, or the protein may lack appropriate post-translational modification. To overcome such limitations, there are a number of strategies that involve using genes from different species, altering constructs, screening for solubility, and utilizing different cellular or cell-free expression systems. The constructs can be altered in numerous ways, such as by the addition of tags, separation of proteins into domains, or the use of gene shuffling methods.

Once the proteins are expressed and purified, it is necessary to form crystals of sufficient quality to collect high-resolution (at least 2.5 Å) data for structure determination. Because crystallization conditions cannot be pre-determined, it is necessary to screen a wide range of conditions (such as pH, salt, protein concentration and co-

factors). Over the past few years, this area has benefited enormously from automation and technologies allowing the use of small sample volumes²⁶. Particularly for proteins and protein complexes with low yields, the ability to screen more conditions at the required protein concentration is critical.

Currently, most biological crystallography experiments are done at synchrotrons, where the brightness (high flux of well-collimated X-rays) and tunability expand the capabilities and throughput enormously. The increase in the amount and diversity of structural data that have been obtained in the past five to ten years has been greatly enhanced by the availability of beamlines and detectors of increasing performance. As the systems have evolved from primitive to 'user-friendly', robotic crystal mounting and alignment systems have also been implemented at beamlines²⁷ to increase the throughput and productivity of these expensive and oversubscribed resources. Once data are obtained, usually in one to several hours on modern third-generation synchrotrons, the analysis of the primary data can also be completed in several hours.

Box 1

Are crystals necessary in electron crystallography?

The fundamental role of crystals within crystallography is that they make it easy to merge the data that are generated by vastly more scattering events than could be tolerated by a single molecule. But because the alignment of high-resolution images of single particles can be done *in silico*, one has to seriously ask whether or not crystals are really needed. Indeed, it now is easier to use single particles than crystals to obtain 3D reconstructions at a resolution of 1–2 nm, as long as the particle has a relative molecular mass of at least 250K–500K.

On the other hand, because the best electron micrographs rarely provide data whose quality is as good as 10% of what physics would allow it to be, it is thought that computational alignment at atomic resolution will require that particles be larger than ~2,000K–4,000K (ref. 43). Furthermore, the number of molecular images that must be merged is approximately 100 times more than would be required if the image quality were nearly perfect. The task of merging data from images of as many as one million individual particles, the number currently used to obtain high-resolution structures of specimens prepared as 2D crystals, is estimated to require at least 10^{17} floating-point operations⁶³. This task would require a full day of dedicated use of a teraflop computer. Improvements in affordable clusters will soon

bring this much computing power into a well-equipped cryo-EM facility, and thus it is likely that computational capacity will keep pace with the projected improvement in speed of data collection. The limiting factor for 2D crystals, on the other hand, is the low rate at which high-resolution images are obtained with highly tilted samples. The steep decline in success of recording images at high tilt angle is the result of some form of specimen charging or beam-induced movement that is still not fully understood.

Although work at atomic resolution with 2D crystals remains an attractive approach for high-profile and otherwise intractable specimens such as tubulin³⁵, a solution to the problem of beam-induced movement must still be found before 2D crystals can be used for work at a pace comparable to that of X-ray crystallography. It is likely that cryo-EM images of single-particle specimens would approach atomic resolution, if we could overcome the same problem that now limits the image quality in highly tilted 2D crystals. Should this type of improvement in performance be realized, accurate alignment of nearly any macromolecular complex could be done *in silico*, and crystals would indeed no longer be needed for crystallography.

Box 2

Experimental methods for structural characterization of assemblies

A variety of methods are available for the experimental determination of macromolecular assembly structure (see Fig. 4)

X-ray crystallography is the most powerful method for structure determination because it is capable of providing an atomic structure of the whole assembly^{22,64}. When suitable crystals and high-resolution crystallographic data are obtained, there is little need for other methods of structure characterization.

Nuclear magnetic resonance (NMR) spectroscopy allows determination of atomic structures of increasingly large subunits and even their complexes^{65–69}. Although NMR analysis is generally not as applicable as X-ray crystallography to protein structures with more than 300 amino acid residues, it can be applied to molecules in solution and is more suitable than X-ray crystallography to study their dynamics and interactions in solution.

Electron crystallography (two-dimensional electron microscopy or 2D EM) and single-particle analysis can reveal the shape and symmetry of an assembly, sometimes at near-atomic resolution, but more frequently at an intermediate resolution⁷⁰. Segmentation of the electron density may lead to an approximate configuration of subunits in a complex⁷¹. Proteins whose structures are already known can then be fitted into these density maps with an accuracy approaching one-tenth the resolution of the EM reconstruction^{47–50}.

Electron tomography is based upon multiple tilted views of the same object⁵⁴. Although it can be used to study the structure of isolated macromolecular assemblies at relatively low resolution, its true potential lies in visualizing the assemblies in an unperturbed cellular context.

Immuno-electron microscopy can be used to determine an approximate position of a protein in the context of an assembly⁷². This task is achieved by using a construct of the protein of interest that binds to a gold-labelled antibody. The relative position of the gold particles is then identified by EM.

Chemical crosslinking with mass spectroscopy can be used to identify binary and higher-order protein contacts⁷³. The approach relies on bi- and tri-functional crosslinking reagents that covalently link proteins interacting with each other. Proteolytic digestion and subsequent mass spectroscopic identification of the crosslinked species reveal their composition. In addition, chemical crosslinking of specific residue types has recently been used to obtain intramolecular distance restraints⁷⁴.

Affinity purification with mass spectroscopy combines purification of protein complexes with identification of their individual components by mass spectroscopy (see reviews in this issue by Aebersold and Mann, page 198, and Fields and co-workers, page 208). During cell lysis, the whole assembly is partially broken into smaller complexes that are then isolated by a variety of methods, such as those relying on fusion

proteins or antibodies as baits for affinity purification. Subunits in these smaller complexes are usually identified by a combination of gel electrophoresis and mass spectroscopy. Examples include the U1 subunit of the yeast and human spliceosome^{75,76}, identification of proteins that interact with the GroEL complex⁷⁷, the sampling of protein interactions in the yeast nuclear-pore complex⁷², and a high-throughput identification of the hundreds of distinct protein complexes in budding yeast^{13,78}.

Fluorescence resonance energy transfer (FRET) occurs when a higher-energy fluorophore stimulates emission by a lower-energy fluorophore that is within ~60 Å of its inducer. It can be applied to monitor protein interactions if one protein is fused to a fluorescence donor and its potential partner to a fluorescence acceptor (see accompanying review by Fields and co-workers). Fluorescence donors and acceptors are usually spectral derivatives of the green fluorescence protein.

Site-directed mutagenesis and a variety of biochemical experiments (for example, footprinting) can reveal which subunits in a complex interact with each other and sometimes what face is involved in the interaction^{79–81}.

Yeast two-hybrid system detects binary protein interactions by activating expression of a reporter gene upon direct binding between the two tested proteins (see review by Fields and co-workers). The approach is based on the modularity of transcription factors that consist of a DNA-binding and an activation domain, each of them fused to two different genes encoding for the proteins whose interaction is tested. If the two expressed fusion proteins are in contact with each other, the two modules of the transcription factor are united, thereby inducing transcription of a set of reporter genes. Expression of reporter genes, in turn, is easily detected by a variety of tests, such as yeast colony colour and ability to grow in deficient media. The method is suitable for high-throughput applications (ref. 11; and see review by Fields and co-workers).

Protein arrays immobilize a variety of 'bait' proteins, such as antibodies and glutathione S-transferase, into an array on a specially treated surface; the array is then probed with sample proteins, resulting in a detection of binary interactions (see review by Fields and co-workers). Messenger RNA expression arrays immobilize stretches of mRNA and are used to measure the concentration of mRNA species in a sample as a function of tissue type, cell cycle and other environmental conditions^{82,83}. Such data sets have been used to detect functionally linked proteins, which include proteins whose expression is co-regulated because they are members of the same assembly, are encoded on the same operon, or belong to the same biochemical pathway⁶.

Increasingly, therefore, structures are solved within hours after data collection begins, although most structures still need a great deal more time for the screening of crystals, full data collection, and the processing and analysis that leads to an accurate high-resolution structure. Nevertheless, as the beamlines become more automated and as higher-level control and processing software is further developed, it is becoming feasible to integrate the data collection, processing and analysis steps — from crystal mounting through structure refinement — to form a 'pipeline' of information for structure determination. The technological advances, such as third-generation synchrotrons and charge-coupled device (CCD)-based detectors, have also been critical for the success of structure determinations of several large complexes and viruses. Crystals from such samples typically have very large unit-cell dimensions and diffract even more weakly than 'ordinary' biomolecular crystals.

Recently, several international efforts have been initiated to determine the structures of at least one member from each domain family,

such that the structures of the remaining protein sequences can be characterized based on their similarity to the known structures^{28,29}. Structural genomics aims to construct a taxonomy of protein structures that will serve as a 'dictionary' for the interpretation of the genomic data. In the United States, the Protein Structure Initiative of the National Institute of General Medical Sciences (NIGMS) has funded nine pilot centres to develop high-throughput pipelines for structure determination³⁰. The NIGMS initiative is paralleled by similar efforts in Europe and Japan. Following the success of the genome sequencing programmes, where the use of automation has been important in the increase of productivity, these structural genomics programmes are currently implementing automation of protein production, crystallization, data collection and analysis.

Although it is legitimate to ask how successful structural genomics will be in terms of structures solved versus targets chosen for cloning, a fair assessment at this point in time is difficult. In the

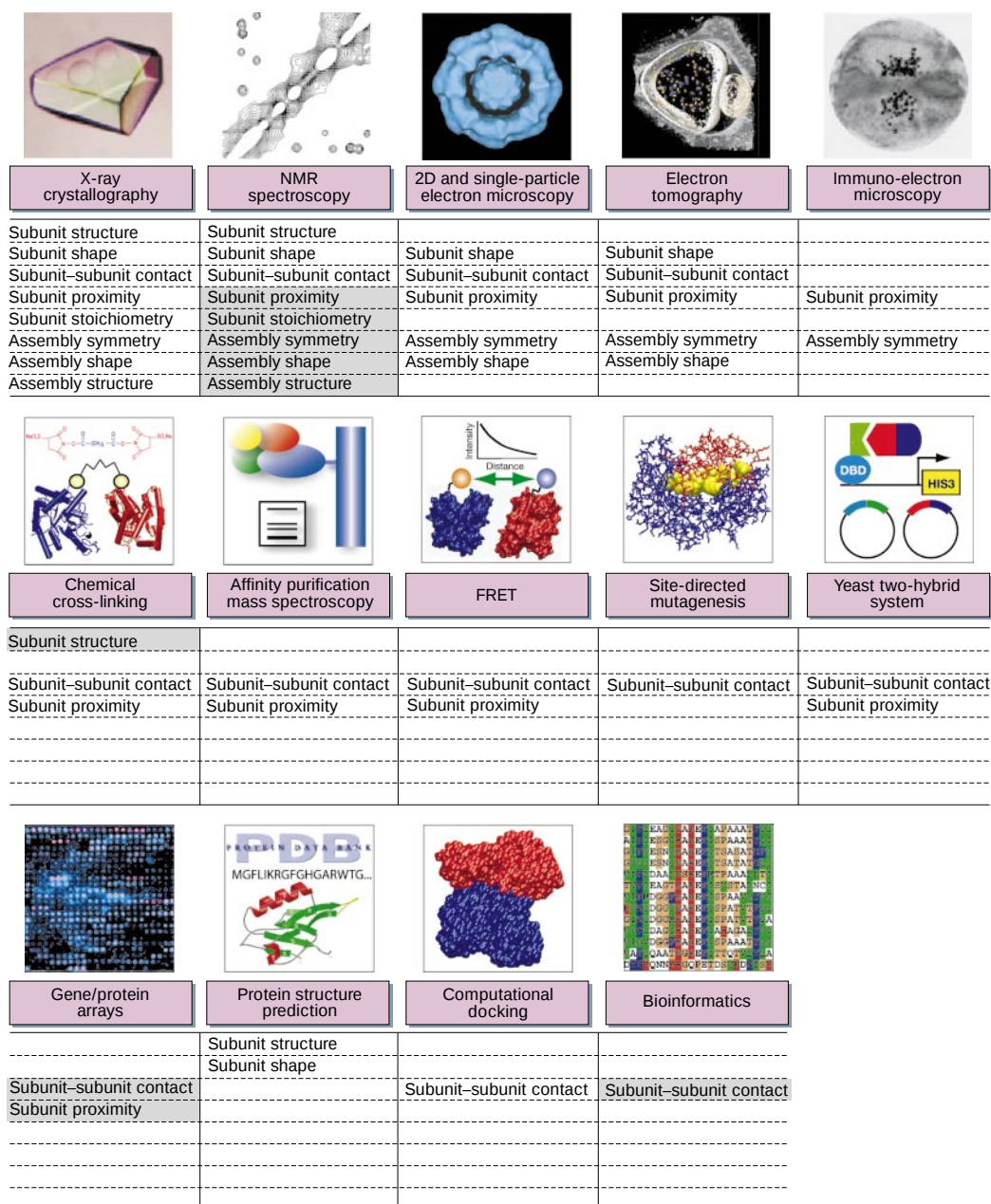


Figure 4 Experimental and theoretical methods that can provide information about a macromolecular assembly structure. The annotations below each of the panels list the aspects of an assembly that might be obtained by the corresponding method. Subunit and assembly structure indicate an atomic or near-atomic resolution at 3 Å or better. Subunit and assembly shape indicate the density or surface envelope at a low resolution of worse than 3 Å. Subunit-subunit contact indicates knowledge about protein pairs that are in

contact with each other, and in some cases about the face that is involved in the contact. Subunit proximity indicates whether two proteins are close to each other relative to the size of the assembly, but not necessarily in direct contact. Subunit stoichiometry indicates the number of subunits of a given type that occur in the assembly. Assembly symmetry indicates the symmetry of the arrangement of the subunits in the assembly. Grey boxes indicate extreme difficulty in obtaining the corresponding information by a given method.

early years, it is first necessary to establish the appropriate infrastructure, and it will take time until this investment pays dividends. Success also depends on the choice of targets; there are easy proteins and families, as well as more difficult ones, such as membrane proteins. Whereas success rates of 1–10% per attempted protein are often quoted, this estimate may be misleadingly pessimistic. Many target families have >10 members, a large number of which are usually attempted in parallel. Therefore, the likelihood that at least one of the targeted family members yields a structure is higher

than 10%. Whatever the timeframe may be, there is no doubt that structural genomics will make a major contribution to the proteomics dictionary of words and phrases. But words or even phrases alone do not make literature.

Using EM images to produce three-dimensional structures

A powerful advantage of EM is the fact that it is possible to treat images of single molecules in the same way as crystalline arrays. The ability to use non-crystalline particles means, in turn, that it is possible to work

Box 3

Theoretical methods for structural characterization of assemblies

Non-experimental methods used to provide information about macromolecular assembly structure include protein structure prediction, computational docking and a variety of bioinformatics techniques (see Fig. 4).

Protein structure prediction can be used to characterize sequences whose structures have not been obtained experimentally⁸⁴. There are two types of methods corresponding to the two distinct sets of principles that guide the behaviour of proteins on vastly different timescales: the laws of physics and the rules of evolution.

The first approach, *de novo* or *ab initio* methods, predicts the structure from sequence alone, without relying on similarity at the fold level between the modelled sequence and any of the known structures⁸⁵. The *de novo* methods assume that the native structure corresponds to the global free-energy minimum accessible during the lifespan of the protein and attempt to find this minimum by an exploration of many conceivable protein conformations. The two key components of *de novo* methods are the procedure for efficiently carrying out the conformational search, and the free-energy function used for evaluating possible conformations.

De novo prediction of protein structure directly from its sequence is becoming increasingly more successful. For roughly 35% of proteins shorter than 150 amino acids that have been examined, one of the five most commonly recurring models generated has sufficient global similarity to the true structure to recognize it in a search of the protein structure database⁸⁶. But the accuracy of even the 'correct' models tends to be only ~4 Å root-mean-square deviation (RMSD) over ~80 residues, too low for problems requiring high-resolution structure information.

The second class of methods of protein structure prediction, including threading and comparative or homology modelling, rely on detectable similarity spanning most of the modelled sequence and at least one known structure⁸⁷. Modelling of a sequence based on known structures consists of four steps: finding known structures related to the sequence to be modelled, aligning the sequence with the related structures, building a model, and assessing the model. The templates for modelling may be found by sequence comparison methods or by sequence–structure threading methods that can sometimes reveal more distant relationships than purely sequence-based methods⁸⁸. In the latter case, fold assignment and alignment are achieved by threading the sequence through each of the

structures in a library of all known folds; each sequence–structure alignment is assessed by the energy of a corresponding coarse model, not by sequence similarity as in sequence comparison methods. Next, given a sequence–structure alignment, comparative model building produces an all-atom model of the sequence.

High-accuracy comparative models are based on more than 50% sequence identity to their templates. They tend to have approximately 1 Å RMS error for the main-chain atoms, which is comparable to the accuracy of a medium-resolution nuclear magnetic resonance structure or a low-resolution X-ray structure. Low-accuracy comparative models are based on less than 30% sequence identity, and tend to contain less than 70% of residues within 3.5 Å of their correct positions. It is currently possible to model domains in ~60% of all known protein sequences⁸⁹. Although the current number of modelled proteins may look impressive given the early stage of structural genomics, usually only one domain per protein is modelled (on the average, proteins have slightly more than two domains) and two-thirds of the models are based on less than 30% sequence identity to the closest template.

Computational docking is based on maximizing the shape and chemical complementarities between a given pair of interacting proteins⁹⁰. Although these methods are generally not yet sufficiently accurate to predict whether two proteins actually interact with each other, they can sometimes correctly identify the interacting surfaces between two known or modelled subunits⁹¹.

Bioinformatics analysis of genomic sequences, multiple sequence alignments and protein structures may indicate the presence and location of protein interaction interfaces. For example, a pair of proteins in a given genome that appear as a fused multidomain protein in another genome indicates a binary interaction between the two proteins in the first genome^{6,92}. Likewise, co-occurrence of two proteins in the same genomic neighbourhood indicates a functional link, especially in prokaryotes⁹³. Similarity between the phylogenetic trees for two families of orthologues also indicates an interaction^{6,94,95}. Correlated mutations resulting in co-variation between alignment positions in two families of proteins are a weak signal that members of the two families may interact with each other⁹⁶. Analyses of multiple sequence alignments and known protein structures, such as the evolutionary trace method⁹⁷, may help in identification of a binding site on a given protein structure. And finally, interactions may be inferred from considerations of protein sequence and structure homology^{98,99}.

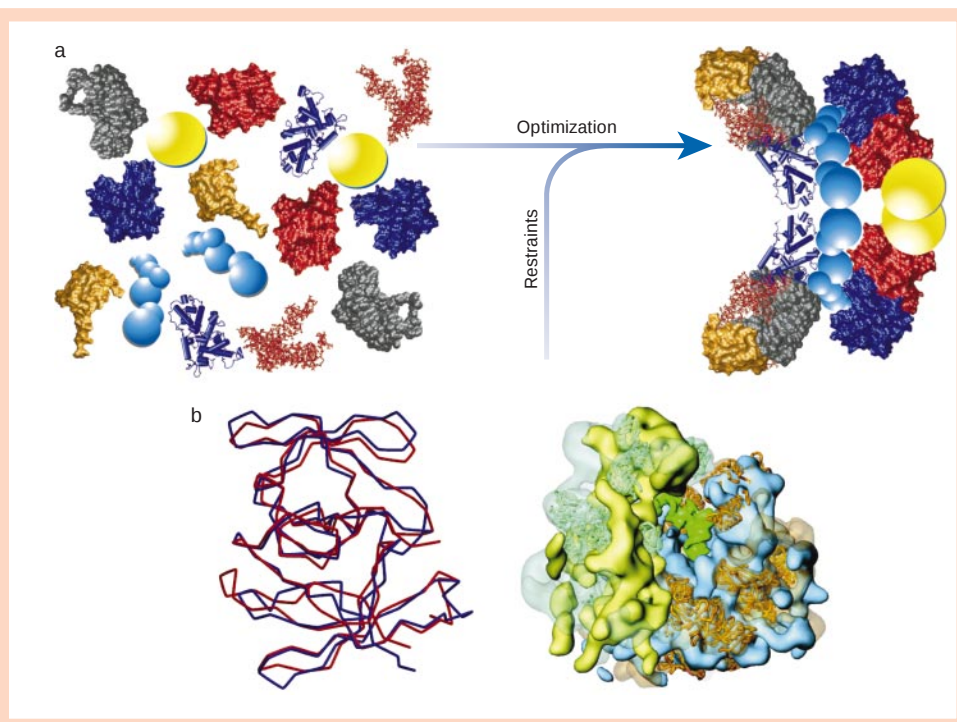
with very small quantities of material, the purity need not be at the standard required for crystallization, and specimen tilting (a bottleneck discussed below) is not needed to collect data for a 3D reconstruction. The electron microscope produces images that represent only 2D projections of the specimen, in which all information about the third dimension of the object has been lost. Nevertheless, the full 3D structure of the object can be reconstructed again if one is able to start with many such projections, each showing the object from a different angle³¹. As a result, the unique contributions that can be made by EM include studies of large, complex assemblies without any requirement for crystallization, and, as will be discussed later, even their visualization within whole cells by electron tomography.

Unfortunately, the electrons in a microscope also represent a beam of ionizing radiation that damages the sample while the image is being formed. As a result, it is necessary to limit the electron exposure to a value that is so low that the images have extremely high levels of 'shot noise' (statistical variation in the number of electrons recorded at each point in the image). Equivalent images of separate molecules must therefore be averaged to reduce the statistical noise that is present in each such image.

If the specimen is one molecule thick with all molecules in the same orientation (as in a 2D crystal), the necessary spatial averaging of images is easy. In fact, 3D reconstructions that have been obtained at a high enough resolution to trace the polypeptide chain have all been produced with the use of 2D crystals^{32–38}. Although only ~100 images of highly tilted crystals are needed to produce such a reconstruction, collection of this amount of experimental data is nevertheless slow because the yield of good images drops to 1% or less of that obtained with untilted specimens. As a result, structural studies with 2D crystals have only seldom been taken to a high enough resolution to allow building an atomic model directly into the 3D reconstruction.

Other specimens may exist in the form of long helices or other particles with very high symmetry (for example, icosahedra). These high-symmetry particles usually do not need to be tilted, as the individual particles are naturally rotated by a random amount relative to one another. The number of protein monomers within one such particle remains relatively small, and thus data from many equivalent particles may still have to be averaged to obtain a reconstruction. In practice, such reconstructions have rarely extended

Figure 5 Hybrid approaches to structure determination of macromolecular complexes. **a**, Scheme illustrating the integration of a diverse set of structures varying in reliability and resolution into a hypothetical hybrid assembly structure. **b**, Hybrid assembly of the 80S ribosome from yeast³⁴. Superposition of a comparative protein structure model for a domain in protein L2 from *Bacillus stearothermophilus* with the actual structure (1RL2) (left). A partial molecular model of the whole yeast ribosome (right) was calculated by fitting atomic rRNA (not shown) and comparative protein structure models (ribbon representation) into the electron density of the 80S ribosomal particle.



beyond about 7–8-Å resolution^{39,40}. Even so, the ability to visualize elements of secondary structure at this resolution makes it easy to fit a previously determined atomic model of protein monomers into the density. The recent docking of the atomic structure of tubulin into the EM density map of a complete microtubule⁴¹ illustrates just how precise this docking can be. This type of docking can then provide accurate images of the protein–protein contacts that lead to the assembly of larger macromolecular machines (Fig. 2).

Because the electron microscope produces images, and not only diffraction intensities, it is possible to determine the positions and relative orientations of randomly distributed, asymmetric macromolecules. The individual images must then be sorted into a large number of distinct classes of views before they can be averaged. This step in the process is illustrated in Fig. 3a, which shows a gallery of 12 different class averages obtained from ice-embedded specimens of *Drosophila melanogaster* tripeptidyl peptidase II (TPP II)⁴². Once a large set of views is in hand, the 3D reconstruction is computed in much the same way as if the average projections had been computed from images of tilted, 2D crystals. As in the example of TPP II that is shown in Fig. 3b, the resulting 3D reconstruction immediately shows how a large, multi-protein complex is assembled from its individual parts. These single particles must be large in size, however, to provide sufficient signal for the alignment at high resolution⁴³. In addition, structure determination by single-particle cryo-EM involves far greater amounts of computation than does structure determination based on 2D crystals or particles with very high internal symmetry (Box 1).

Although the capabilities of single-particle cryo-EM are powerful, the method still remains slow compared to other structure-determination technologies, such as X-ray crystallography or NMR spectroscopy. Completion of a structure at the modest resolution of ~2 nm currently may require a month or more for data collection and perhaps another month for data processing. If the goal is to obtain a density map in which features of secondary structure are clearly visible, data collection may extend over several months. A further drawback of cryo-EM is the fact that data collection remains a specialist craft that requires many months, even years of training, before one is able to take full advantage of the high performance of modern electron microscopes.

But it is not necessary for data collection to take as long as it currently does, or to be so dependent on the scientist having a high level of acquired technical expertise. Instead, recording a large number of particle images that are invisible to the human eye on the viewing screen involves blindly following a prescribed sequence of repetitive operations. In principle, such a task is better suited for a computer than a human operator. Indeed, automated implementations of single-particle data-collection operations have recently been published^{44,45}. The next frontier where work has already begun includes automation of the steps in which images of individual particles are selected within digitized micrographs and the data are merged into a 3D reconstruction of the particle. In one recent demonstration, for example, data were collected and a 3D reconstruction was obtained for the tobacco mosaic virus particle at a resolution of ~1 nm in a period of less than 24 hours⁴⁶. Further development of automated data collection and analysis promises to reduce the turnaround time for producing 3D density maps of large, macromolecular particles from months or years to days or weeks.

The 3D reconstructions obtained by cryo-EM are likely to be used primarily for docking (that is, assembling) atomic-resolution models of component macromolecules into the 3D densities of intact complexes. When the resolution of the density map is high enough to see helices and regions of β -sheet, the docking can be done precisely and with little ambiguity. At lower resolution, however, the docking must be performed with caution, and researchers continue to develop quantitative criteria that can guide the operation^{47–50}. It is therefore fortunate that the throughput of cryo-EM should soon become well matched to the combined throughput of X-ray crystallography and NMR spectroscopy, which are the primary sources of the structures of the individual components. In turn, atomic models of the various assembled components can then be used to interpret each of the recognizable densities that are visualized within whole-cell tomograms.

Hybrid approaches to structure determination

X-ray crystallography may provide high-resolution structures of large complexes, if they can be purified in sufficient quantities and crystallized. Single-particle EM can provide medium-resolution

structures (~1 nm) of complexes even if only small amounts of material are available and can tolerate some sample heterogeneity. Even so, these 'direct' methods are surely not capable of characterizing the myriads of stable complexes that exist in a cell. In addition, most of the transient complexes cannot be addressed at all with these approaches. Therefore, there is a great need for hybrid methods where both high throughput and highest possible resolution are achieved by integrating information from different sources. This integration should be performed in an objective manner, such that it is reproducible by any expert.

The hybrid assembly of a complex needs to reflect spatial restraints of varying accuracy and resolution that originate from vastly different experiments and theoretical considerations (Fig. 4, and Boxes 2 and 3). To this end, it is useful to express structure determination as an optimization problem. In this view, 3D models that are consistent with the input information are calculated by optimizing a scoring function. The three components of this approach are: representation of an assembly; a scoring function consisting of individual spatial restraints; and optimization of the scoring function to obtain the models. Figure 5a illustrates how the subunits of a hypothetical complex (left) can be assembled through optimization with respect to restraints from a variety of methods to obtain the final assembly model (right). Each subunit in an assembly can be represented by a set of points that depend on what is known about the subunit. If an experimentally determined structure of a protein is available or a comparative protein structure model can be calculated, each atom can be represented by its own point. If protein domains can be assigned based on biochemical characterization or bioinformatics analysis (for example, by scanning against a sequence database of domains or by prediction of transmembrane spanning domains), a single point represents each domain. Otherwise, a single point can represent the whole subunit.

The most important aspect of the calculation is to accurately capture all of the existing experimental and theoretical information about the structure of a modelled assembly. For example, the shape, density and symmetry of a complex may be derived from EM; upper distance bounds on residues from different subunits may be obtained from X-ray crystallography or NMR spectroscopy and chemical crosslinking; and protein–protein contact restraints may be obtained from immuno-purification with mass spectroscopy and bioinformatics analysis of an alignment of homologous sequences. An 'ensemble' of models that minimize violations of the input restraints can be obtained by optimizing the scoring function, relying on an optimization method such as simulated annealing with molecular dynamics applied in Cartesian space. Because the optimization is likely to be stochastic, a large number of models need to be calculated and assessed. Examples of predicting assembly structures through satisfaction of varied spatial restraints include the *Escherichia coli* 30S ribosomal subunit⁵¹ and the yeast exosome⁵².

A sample study that illustrates some of the points made above is the hybrid assembly of the 80S ribosome (Fig. 5b). A partial molecular model of the whole yeast ribosome was calculated by fitting atomic ribosomal RNA and comparative protein structure models into the electron density of the 80S ribosomal particle, obtained by EM at 15-Å resolution⁵³. Most of the models for 40 out of the 75 ribosomal proteins were based on approximately 30% sequence identity to their template structures. Typical accuracy of a comparative model in this range of sequence similarity is indicated by a comparison of a model for a domain in protein L2 from *Bacillus stearothermophilus* with the actual structure. The fitting of the subunits into the electron density was made possible by the atomic structures of the whole small and large ribosomal subunits from archaea.

Visualizing complexes using electron tomography

Electron tomography is by no means a new imaging technology, but it has only recently gathered momentum^{54,55} (Fig. 6). With the advent of

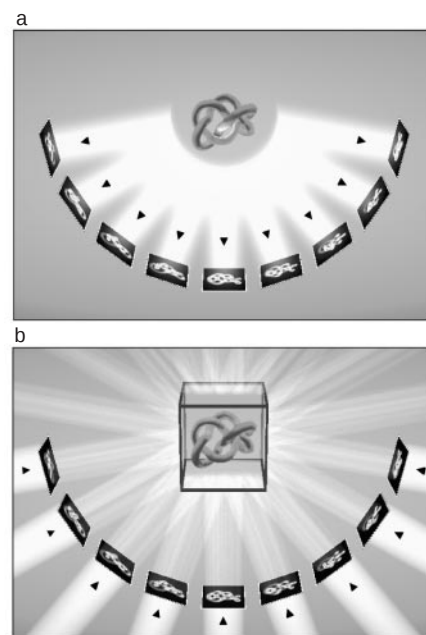


Figure 6 Principle of electron tomography. **a**, Schematic representation of data acquisition. A flexible rope knot represents the object, emphasizing that electron tomography can retrieve 3D information from structures with individual topologies. A set of projection images is recorded on a charge-coupled device camera while the object is tilted incrementally around an axis perpendicular to the electron beam. Owing to the limited accuracy ('eucentricity') of the tilting device, the specimen has to be re-centred and refocused at each tilt angle. Automated procedures have been developed to perform this task with negligible exposure of the object to the electron beam⁶². **b**, The back-projection method explains the principle of the 3D reconstruction in an intuitive manner. For each projection, a back-projection body is calculated, and the sum of all projection bodies yields the density distribution of the original object — the tomogram. To compensate for the fact that high-resolution features change more rapidly with tilt angle than do low-resolution features, an appropriate weighting function has to be applied to the data in the 2D images before calculating the reconstruction. The quality of a tomogram depends critically on covering as wide a tilt range as possible (typically $\pm 70^\circ$), with tilt increments as small as possible (1 to 3°). However, each additional exposure to the beam increases the amount of radiation damage and the cumulative dose must not exceed a tolerable limit.

computer-controlled electron microscopes and the automation of elaborate image acquisition procedures, it became possible to obtain molecular-resolution tomograms of structures as large and complex as whole prokaryotic cells or thin eukaryotic cells embedded in amorphous ice⁵⁶. Non-invasive imaging of whole, vitrified cells is where electron tomography can make a unique contribution and will probably have the greatest impact. The emerging picture of the cell is one of a giant supra-molecular assembly; but on the nanoscale, the cytoplasm is mostly an uncharted territory. Just as high-resolution 3D structures of macromolecules provide valuable insights into their working, a better understanding of cellular functions will arise from the ability to visualize macromolecules in an unperturbed cellular context.

Tomograms of cells at molecular resolution are essentially 3D images of the cell's entire proteome. They reveal information about the spatial relationships of macromolecules in the cytoplasm, the 'interactome'. But exploitation of this information is confronted with two problems. Cryo-tomograms are contaminated by substantial residual noise and distorted by missing data resulting from the restricted tilt range. Moreover, the cytoplasm is very densely

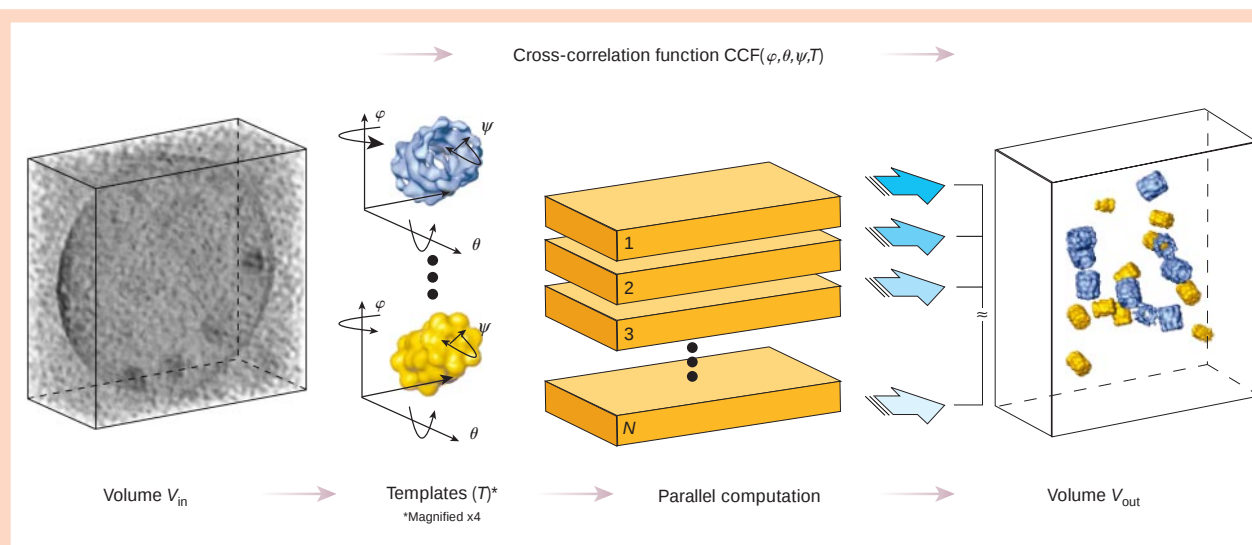


Figure 7 Mapping the spatial distribution of complexes and their interactions within cells. A molecular-resolution tomogram of a cell (V_{in}) is essentially a 3D image of the cell's entire proteome. Residual noise and molecular crowding hamper visualization of the information that is present in this tomogram. As a result, 3D pattern recognition must be used to 'mine' this information. One approach that has been demonstrated to be effective is template matching. Templates of the macromolecular complexes (or even parts thereof) that are of interest must first be obtained by techniques such as cryo-EM or hybrid X-ray/NMR and cryo-EM reconstruction. These templates T

(magnified $\times 4$ in this figure) are used to search for matching structures by cross-correlation, and the result is refined by multivariate statistical analysis. Because both the positions of the complexes and their orientations are initially unknown, V_{in} must be scanned for all possible orientations of each of the templates. The result (V_{out}) shows the positions and orientations of the complexes in the cell. In principle, it should be possible to chart the cellular 'interactome' — the spatial relationships of all major complexes of a cell — by this approach.

populated ('molecular crowding'), with molecules literally touching each other³⁷. Under these conditions, segmentation and feature extraction based on visual inspection is usually impossible, except for some easily recognizable features, such as membranes and the cytoskeleton.

Nevertheless, pattern-recognition techniques can be used, in one guise or another, to detect and identify specific molecules⁵⁸. Provided that a high- or medium-resolution structure of the molecule of interest is available, it can be used as a template to perform a systematic search of the reconstructed volume for matching structures (Fig. 7). Such a molecular signature-based approach, while computationally demanding, can be efficiently parallelized. Once the spatial coordinates of a complex in a cell have been determined, sub-tomograms that encompass the complex and its neighbourhood can be extracted for further analysis and averaging. Multivariate statistical analysis of such sub-tomograms can be used to explore variations in their functional environment⁵⁹.

The feasibility of template matching has been demonstrated with 'phantom cells' (lipid vesicles filled with macromolecules), which provide a realistic experimental scenario and facilitate an assessment of the fidelity of the approach. With the current (non-isotropic) resolution of 4–5 nm, one can address only larger ($M_r > 400K$) complexes in a cellular context. To widen the scope of cellular tomography, it will be necessary to improve the resolution. Theoretical considerations⁶⁰ and ongoing instrumental improvements (such as liquid helium versus liquid nitrogen temperature, improved detectors and dual-axis tilting) make a resolution near 2 nm a realistic goal⁶¹.

Perspectives

The possibility seems now assured of assembling a structural picture that can be 'zoomed' continuously from the details of atomic models all the way up to the full complexity of an intact cell. Structural genomics will bring us closer to a comprehensive dictionary of proteins in the foreseeable future, while EM techniques and hybrid approaches will allow us to assemble proteins as words into meaningful sentences. A comprehensive description of large complexes will

generally require the use of a number of experimental models (Box 2), underpinned by a variety of theoretical approaches (Box 3) to maximize efficiency, completeness, accuracy and resolution of the experimental determination of assembly composition and structure. In conjunction with the non-invasive 3D imaging of whole cells, these approaches might ultimately enable us to read the molecular book of the cell. □

doi:10.1038/nature01513

- Alberts, B. The cell as a collection of protein machines — preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).
- Baumeister, W. & Steven, A. C. Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.* **25**, 624–631 (2000).
- Sali, A. & Kuriyan, J. Challenges at the frontiers of structural biology. *Trends Biochem. Sci.* **24**, M20–M24 (1999).
- Orengo, C. A. et al. The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* **2**, 11–21 (2002).
- Govindarajan, S., Recabarren, R. & Goldstein, R. A. Estimating the total number of protein folds. *Proteins* **35**, 408–414 (1999).
- Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
- Mewes, H. W. et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).
- Costanzo, M. C. et al. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* **29**, 75–79 (2001).
- von Mering, C. et al. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
- Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574 (2001).
- Uetz, P. et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
- Aloy, P. & Russell, R. B. Potential artefacts in protein–interaction networks. *FEBS Lett.* **530**, 253–254 (2002).
- Gavin, A. C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
- Aloy, P. & Russell, R. B. The third dimension for protein interactions and complexes. *Trends Biochem. Sci.* **27**, 633–638 (2002).
- Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein–protein interactions. *Genome Res.* **2**, 37–46 (2002).
- Ge, H., Liu, Z., Church, G. M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **4**, 482–486 (2001).
- Edwards, A. M. et al. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **10**, 529–536 (2002).

18. Kumar, A. & Snyder, M. Protein complexes take the bait. *Nature* **415**, 123–124 (2002).
19. Abbott, A. The society of proteins. *Nature* **417**, 894–896 (2002).
20. Westbrook, J. *et al.* The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* **30**, 245–248 (2002).
21. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. *Science* **292**, 1863–1876 (2001).
22. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
23. Harms, J. *et al.* High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **107**, 679–688 (2001).
24. Wimberly, B. T. *et al.* Structure of the 30S ribosomal subunit. *Nature* **407**, 327–339 (2000).
25. Yusupov, M. M. *et al.* Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883–896 (2001).
26. Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. Automation of X-ray crystallography. *Nature Struct. Biol.* **7**, 973–977 (2000).
27. Snell, G. *et al.* Automatic sample mounting and alignment system for biological crystallography. *J. Synchrotron Radiat.* (in press).
28. Burley, S. K. *et al.* Structural genomics: beyond the Human Genome Project. *Nature Genet.* **23**, 151–157 (1999).
29. Vitkup, D., Melamud, E., Moul, J. & Sander, C. Completeness in structural genomics. *Nature Struct. Biol.* **8**, 559–566 (2001).
30. Structural genomics. *Nature Struct. Biol.* **7**(Suppl.), 927–994 (2000).
31. Frank, J. *Three-dimensional Electron Microscopy of Macromolecular Assemblies* (Academic, London, 1996).
32. Henderson, R., Baldwin, J. M. & Ceska, T. A. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **213**, 899–929 (1990).
33. Kuhlbrandt, W., Wang, D. N. & Fujiyoshi, Y. Atomic model of plant light-harvesting complex by electron crystallography. *Nature* **367**, 614–621 (1994).
34. Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J. Mol. Biol.* **259**, 393–421 (1996).
35. Nogales, E., Wolf, S. G. & Downing, K. H. Structure of the $\alpha\beta$ tubulin dimer by electron crystallography. *Nature* **391**, 199–203 (1998).
36. Mitsuoka, K. *et al.* The structure of bacteriorhodopsin at 3.0 Å resolution based on electron crystallography: implication of the charge distribution. *J. Mol. Biol.* **286**, 861–882 (1999).
37. Murata, K. *et al.* Structural determinants of water permeation through aquaporin-1. *Nature* **407**, 599–605 (2000).
38. Lowe, J., Li, H., Downing, K. H. & Nogales, E. Refined structure of $\alpha\beta$ -tubulin at 3.5 Å resolution. *J. Mol. Biol.* **313**, 1045–1057 (2001).
39. Conway, J. F. *et al.* Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. *Nature* **386**, 91–94 (1997).
40. Bottcher, B., Wynne, S. A. & Crowther, R. A. Determination of the fold of the core protein of hepatitis B virus by cryo-electron microscopy. *Nature* **386**, 88–91 (1997).
41. Li, H. L., DeRosier, D. J., Nicholson, W. V., Nogales, E. & Downing, K. H. Microtubule structure at 8 Å resolution. *Structure* **10**, 1317–1328 (2002).
42. Rockel, B., Peters, J., Kuhlmoorgen, B., Glaeser, R. M. & Baumeister, W. A giant protease with a twist: the TPP II complex from *Drosophila* studied by electron microscopy. *EMBO J.* **21**, 5979–5984 (2002).
43. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193 (1995).
44. Carragher, B. *et al.* Leginon: an automated system for acquisition of images from vitreous ice specimens. *J. Struct. Biol.* **132**, 33–45 (2000).
45. Zhang, P. J., Beatty, A., Milne, J. L. S. & Subramaniam, S. Automated data collection with a Tecnai 12 electron microscope: applications for molecular imaging by cryomicroscopy. *J. Struct. Biol.* **135**, 251–261 (2001).
46. Zhu, Y. X., Carragher, B., Kriegman, D. J., Milligan, R. A. & Potter, C. S. Automated identification of filaments in cryoelectron microscopy images. *J. Struct. Biol.* **135**, 302–312 (2001).
47. Rossmann, M. G., Bernal, R. & Pletnev, S. V. Combining electron microscopic with X-ray crystallographic structures. *J. Struct. Biol.* **136**, 190–200 (2001).
48. Wriggers, W. & Birmanns, S. Using Situs for flexible and rigid-body fitting of multiresolution single-molecule data. *J. Struct. Biol.* **133**, 193–202 (2001).
49. Volkman, N. & Hanein, D. Quantitative fitting of atomic models into observed densities derived by electron microscopy. *J. Struct. Biol.* **125**, 176–184 (1999).
50. Chacon, P. & Wriggers, W. Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.* **317**, 375–384 (2002).
51. Malhotra, A., Tan, R. K. & Harvey, S. C. Prediction of the three-dimensional structure of *Escherichia coli* 30S ribosomal subunit: a molecular mechanics approach. *Proc. Natl Acad. Sci. USA* **87**, 1950–1954 (1990).
52. Aloy, P. *et al.* A complex prediction: three-dimensional model of the yeast exosome. *EMBO Rep.* **3**, 628–635 (2002).
53. Spahn, C. M. *et al.* Structure of the 80S ribosome from *Saccharomyces cerevisiae*—tRNA-ribosome and subunit-subunit interactions. *Cell* **107**, 373–386 (2001).
54. Baumeister, W. Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Curr. Opin. Struct. Biol.* **12**, 679–684 (2002).
55. Baumeister, W., Grimm, R. & Walz, J. Electron tomography of molecules and cells. *Trends Cell Biol.* **9**, 81–85 (1999).
56. Medalia, O. *et al.* Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **298**, 1209–1213 (2002).
57. Grunewald, K., Medalia, O., Gross, A., Steven, A. & Baumeister, W. Prospects of electron cryotomography to visualize macromolecular complexes inside cellular compartments: implications of crowding. *Biophys. Chem.* (in press).
58. Bohm, J. *et al.* Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Natl Acad. Sci. USA* **97**, 14245–14250 (2000).
59. Frangakis, A. S. *et al.* Identification of macromolecular complexes in electron cryotomograms of phantom cells. *Proc. Natl Acad. Sci. USA* **99**, 14153–14158 (2002).
60. Grimm, R. *et al.* Electron tomography of ice-embedded prokaryotic cells. *Biophys. J.* **74**, 1031–1042 (1998).
61. Plitzko, J. *et al.* In vivo veritas: electron cryotomography of cells. *Trends Biotechnol.* **20**, S40–S44 (2002).
62. Koster, A. J. *et al.* Perspectives of molecular and cellular electron tomography. *J. Struct. Biol.* **120**, 276–308 (1997).
63. Glaeser, R. M. Electron crystallography: present excitement, a nod to the past, anticipating the future. *J. Struct. Biol.* **128**, 3–14 (1999).
64. Zhang, G. Y. *et al.* Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell* **98**, 811–824 (1999).
65. Fiaux, J., Bertelsen, E. B., Horwich, A. L. & Wuthrich, K. NMR analysis of a 900K GroEL–GroES complex. *Nature* **418**, 207–211 (2002).
66. Yee, A. *et al.* An NMR approach to structural proteomics. *Proc. Natl Acad. Sci. USA* **99**, 1825–1830 (2002).
67. Fushman, D., Xu, R. & Cowburn, D. Direct determination of changes of interdomain orientation on ligation: use of the orientational dependence of ^{15}N NMR relaxation in Abl SH(32). *Biochemistry* **38**, 10225–10230 (1999).
68. Nakanishi, T. *et al.* Determination of the interface of a large protein complex by transferred cross-saturation measurements. *J. Mol. Biol.* **318**, 245–249 (2002).
69. Pellecchia, M., Sem, D. S. & Wuthrich, K. NMR in drug discovery. *Nature Rev. Drug Discov.* **1**, 211–219 (2002).
70. Frank, J. Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 303–319 (2002).
71. Volkman, N. A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. *J. Struct. Biol.* **138**, 123–129 (2002).
72. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148**, 635–651 (2000).
73. Rappsilber, J., Siniosoglou, S., Hurt, E. C. & Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.* **72**, 267–275 (2000).
74. Young, M. M. *et al.* High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl Acad. Sci. USA* **97**, 5802–5806 (2000).
75. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl Acad. Sci. USA* **94**, 385–390 (1997).
76. Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20**, 46–50 (1998).
77. Houry, W. A., Frishman, D., Eckerskorn, C., Lottspeich, F. & Hartl, F. U. Identification of *in vivo* substrates of the chaperonin GroEL. *Nature* **402**, 147–154 (1999).
78. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
79. Miras, I., Schaeffer, F., Beguin, P. & Alzari, P. M. Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of *Clostridium thermocellum* CipA. *Biochemistry* **41**, 2115–2119 (2002).
80. Wells, J. A. Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol.* **202**, 390–411 (1991).
81. Jin, L., Cohen, F. E. & Wells, J. A. Structure from function: screening structural models with functional data. *Proc. Natl Acad. Sci. USA* **91**, 113–117 (1994).
82. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
83. Lockhart, D. J. & Winzler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
84. Baker, D. & Salí, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
85. Bonneau, R. & Baker, D. *Ab initio* protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173–189 (2001).
86. Bonneau, R. *et al.* De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **322**, 65–78 (2002).
87. Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
88. Domingues, F. S., Lackner, P., Andreeva, A. & Sippl, M. J. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.* **297**, 1003–1013 (2000).
89. Pieper, U., Eswar, N., Stuart, A. C., Ilyin, V. A. & Salí, A. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **30**, 255–259 (2002).
90. Smith, G. R. & Sternberg, M. J. E. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12**, 28–35 (2002).
91. Strynadka, N. C. J. *et al.* Molecular docking programs successfully predict the binding of a β -lactamase inhibitory protein to TEM-1 β -lactamase. *Nature Struct. Biol.* **3**, 233–239 (1996).
92. Enright, A. J., Iliopoulos, I., Kyriades, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
93. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96**, 2896–2901 (1999).
94. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293 (2000).
95. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14**, 609–614 (2001).
96. Pazos, F. & Valencia, A. *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219–227 (2002).
97. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).
98. Lappe, M., Park, J., Niggemann, O. & Holm, L. Generating protein interaction maps from incomplete data: application to fold assignment. *Bioinformatics* **17**, S149–S156 (2001).
99. Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA* **99**, 5896–5901 (2002).

Acknowledgements We thank N. Eswar for preparing the histograms in Fig. 1, M. Simon, B. Jap and H. Noller for permission to use the structural images in Fig. 1, K. H. Downing for Fig. 2, B. Rockel for Fig. 3, and F. Alber for Figs 4 and 5a. We are also grateful to P. Bjorkman and H. Moss for commenting on the manuscript. This work has been supported in part by NIH grants (to A.S., R.M.G. and T.E.), the Agouron Institute (T.E.) and a Max-Planck Research Award (W.B.).

Disease proteomics

Sam Hanash

Department of Pediatrics, University of Michigan, 1150 West Medical Center Drive, Ann Arbor, Michigan 48109, USA
(e-mail: shanash@umich.edu)

The sequencing of the human genome and that of numerous pathogens has opened the door for proteomics by providing a sequence-based framework for mining proteomes. As a result, there is intense interest in applying proteomics to foster a better understanding of disease processes, develop new biomarkers for diagnosis and early detection of disease, and accelerate drug development. This interest creates numerous opportunities as well as challenges to meet the needs for high sensitivity and high throughput required for disease-related investigations.

Despite tremendous advances in our understanding of the molecular basis of diseases such as cancer, substantial gaps remain both in our understanding of disease pathogenesis and in the development of effective strategies for early diagnosis and for treatment. The current interest in proteomics is due in part to the prospects that a proteomic approach to disease investigations will overcome some of the limitations of other approaches¹. The opportunities as well as the challenges facing disease proteomics are formidable. Particularly promising areas of research include: delineation of altered protein expression, not only at the whole-cell or tissue levels, but also in subcellular structures, in protein complexes and in biological fluids; the development of novel biomarkers for diagnosis and early detection of disease; and the identification of new targets for therapeutics and the potential for accelerating drug development through more effective strategies to evaluate therapeutic effect and toxicity.

The dynamic nature of the proteome of a cell or a tissue provides ample justification for studying gene expression in disease directly at the proteomic level. But capturing this dynamic state represents a technological challenge. Undoubtedly, tackling the numerous facets of disease proteomics requires implementation of multiple strategies and technology platforms.

Proteome profiling technologies are currently evolving in a manner that emphasizes the need for sensitivity and throughput. No one technology is likely to emerge that will meet the needs of all types of proteomics-based investigations, from expression proteomics to functional proteomics, particularly as they relate to disease.

The use of two-dimensional gels

During the early years of proteomics and until relatively recently, profiling of protein expression in disease relied primarily on the use of two-dimensional polyacrylamide gel electrophoresis (2D PAGE), which was later combined with mass spectrometry². Most studies of this nature followed an approach in which a cocktail was used to solubilize the protein contents of an entire cell population, tissue or biological fluid, followed by separation of the protein contents of the lysate using 2D gels and visualization of the separated proteins using silver staining. It became clear that such an approach allows only a limited display of protein content that consisted of relatively abundant proteins. Nevertheless, profiling of disease tissues

using this approach has had some utility. For example, it was demonstrated long before the use of DNA microarrays that leukaemias could be classified into their different subtypes using 2D PAGE³.

Numerous other studies have also identified disease-related changes in protein expression, primarily using 2D PAGE and mass spectrometry. One such example is provided by studies of heart disease⁴, the spectrum of which encompasses a broad set of pathological conditions, some with acute onset of severe disease and others with slow, chronic progression. To assist in data gathering and mining, online 2D gel-derived databases of protein expression in the myocardium for human and other species were constructed^{5,6}. These databases have allowed investigators to compare data and establish reference standards. Relevant findings have emerged from studies of changes in myocardial proteins associated with human heart failure as well as from studies of animal models of heart failure and of isolated rat myocytes⁴. Although only a small proportion of the proteome has been analysed, pronounced changes in the composition of the cardiac proteome have been found, affecting proteins with diverse functions. Altered overall levels of specific proteins or altered post-translational modifications of proteins such as myosin light chain 2 have been reported in the failing heart⁷. And protein expression studies have uncovered proteins that exhibited new disease-related post-translational modifications with predicted functional relevance^{8–11}. This level of progress typifies that made in profiling disease tissue in a wide variety of diseases.

An important development in 2D PAGE is the use of immobilized pH gradients (IPGs) in which the pH gradient is fixed within the acrylamide matrix. IPGs also allow production of gels that cover a defined pH range, from wide to narrow^{12,13}. A variation on this theme is the use of so-called 'zoom gels' in which the protein contents of an individual sample are first fractionated into narrow pH ranges under low resolution, and then each fraction undergoes high-resolution separation by 2D PAGE¹⁴. For example, the pre-fractionation of serum using this approach has enhanced the ability to detect low-abundance and potentially new circulating disease-marker proteins¹⁴. Yet another innovation in 2D gels is the use of differential in-gel electrophoresis (DIGE), in which two pools of proteins are labelled with different fluorescent dyes¹⁵. The labelled proteins are mixed and separated in the same 2D gel. In one study¹⁶, 2D DIGE was applied to quantify the differences in protein expression between oesophageal carcinoma cells and normal

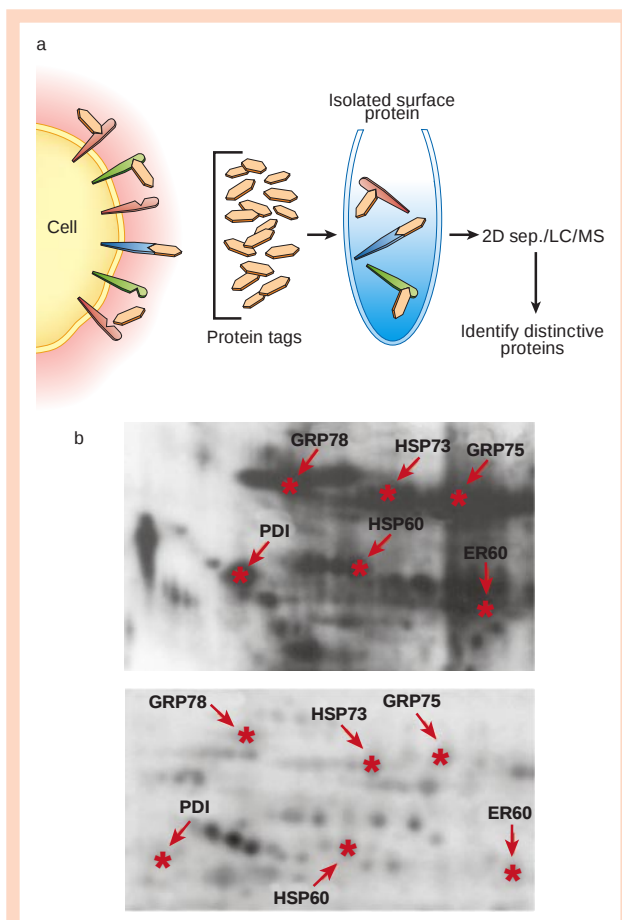


Figure 1 Affinity capture of surface-membrane proteins. Biotinylation reagents provide a 'tag' that transforms poorly detectable surface-membrane proteins into probes that can be recognized by a labelled detection reagent. **a**, Intact cells are tagged using lipid-insoluble biotin reagents. Tagged proteins are captured after cell lysis using avidin columns and subsequently eluted. Following a separation step using 2D gels or other means, tagged proteins are detected with a labelled avidin conjugate. Individual proteins are identified by mass spectrometry. **b**, Close-up section of a 2D pattern in which biotinylated proteins are selectively visualized (top) in contrast with the pattern of the same whole-cell lysate visualized by silver staining (bottom). Several selectively visualized proteins were found to have chaperone functions¹⁹.

epithelial cells. A large number of proteins were found to be either upregulated or downregulated in cancer cells.

Sample fractionation prior to analysis

Some of the main challenges facing expression proteomics, be it using 2D PAGE or any other approach, include the great dynamic range of protein abundance and a wide range of protein properties, including mass, isoelectric point, extent of hydrophobicity and post-translational modifications. Reducing sample complexity prior to analysis — for example, by analysing protein subsets and subcellular organelles separately — improves the reach of 2D gels or other separation techniques for the quantitative analysis of low-abundance proteins. An elegant demonstration of the power of sub-proteome analysis is illustrated in studies of phagosomes¹⁷; these have led to the identification of over 250 proteins from this organelle and the demonstration that phagosomes are formed by direct association and fusion of endoplasmic reticulum to the plasma membrane during early phagocytosis.

The isolation of sub-proteomes may be combined with protein tagging to further enhance sensitivity, as in the case of surface-

membrane proteins, a compartment rich in diagnostic and therapeutic targets. Protein tagging technologies are currently being implemented for the comprehensive analysis of the cell-surface proteome (Fig. 1). A surface-protein biotinylation strategy, coupled with the use of mass spectrometry, was applied to the gastric pathogen *Helicobacter pylori*, leading to the identification of new surface-membrane proteins¹⁸. This strategy has also led to the detection and identification of many new proteins on the surface of cancer cells (Fig. 1)¹⁹.

Beyond 2D gels for disease expression profiling

Even with all the improvements that could be introduced, 2D gels will probably remain a rather low-throughput approach that requires a relatively large amount of sample. The latter is particularly problematic for clinical samples, as such samples are generally procured in limited amounts. Furthermore, tissue heterogeneity complicates the analysis of clinical samples. Various tissue microdissection approaches are beneficial to reduce heterogeneity, but they further reduce the amount of sample available. In particular, the use of laser-capture microdissection, which allows defined cell types to be isolated from tissues, yields amounts of proteins that are difficult to reconcile with the need for greater amounts for 2D gels¹.

Undoubtedly, various non-gel-based schemas that rely on liquid-based separations of proteins or peptides, with or without tagging, will have utility for disease proteomics, particularly given their potential for automation. Additionally, advances in microfluidic technology will likely allow automated separation of proteins in complex lysates using much reduced sample amounts. Microfluidic systems already have been integrated with mass spectrometry for protein digestion and identification²⁰.

Non-separation-based strategies, including direct profiling using mass spectrometry or the use of protein microarrays, are important developments. Mass spectrometry has been applied to the *in situ* proteomic analysis of tissues, an approach that allows imaging of protein expression in normal and disease tissues²¹. By this method, frozen tissue is sliced and sections are applied on a matrix-assisted laser desorption/ionization (MALDI) plate and analysed at regular spatial intervals. The mass spectra obtained at different intervals are compared, yielding a spatial distribution of individual masses across the tissue section (Fig. 2). Mass profiles of tissue sections obtained from normal and disease tissues may be compared to detect altered protein expression. Tumour analyses using this approach have uncovered differences in protein expression between normal and tumour tissues that may have specificity for different tumour types²¹.

DNA and protein microarrays in disease investigation

Cancer profiling using DNA microarrays

Profiling gene expression using DNA arrays has had a tremendous impact on biomedical research. Disease-related applications of DNA microarrays include uncovering unsuspected associations between genes and specific clinical features of disease that are helping devise new molecular-based classifications of disease. In relation to cancer, most published studies of tumour analysis using DNA microarrays have examined pathologically homogeneous sets of tumours to identify clinically relevant subtypes (for example, responders versus non-responders), pathologically distinct subtypes of tumours of the same lineage to identify molecular correlates (for example, high-stage versus low-stage tumours), or tumours of different lineages to identify molecular signatures for each lineage.

Published studies of breast cancer illustrate the potential contribution of DNA microarrays to uncover new disease subtypes. In one study, tumours could be classified into a basal epithelial-like group, an ErbB2-overexpressing group and a normal breast-like group²². In a later study, survival analyses on a sub-cohort of patients with locally advanced breast cancer showed significantly different outcomes for patients belonging to the various groups, despite uniform treatment²³. In an independent study of 38 invasive breast cancers,

striking molecular differences between ductal carcinoma specimens were uncovered that led to a suggested new classification for oestrogen-receptor (ER)-negative breast cancer²⁴. Similarly, a study of 58 node-negative breast carcinomas discordant for ER status also uncovered a list of genes that discriminated tumours according to ER status²⁵. More recently, gene expression profiling was found to be a more powerful predictor of disease outcome in young patients with cancer than clinical- and histological-based classifications²⁶.

DNA versus protein microarrays

The DNA microarray studies described above, as well as numerous others in the literature, indicate the great utility of DNA microarrays for uncovering patterns of gene expression that are clinically informative. An important challenge for microarray analysis of disease tissues and cells is to understand at a mechanistic level the significance of associations observed between subsets of genes and clinical features of disease. Another challenge is to identify the smallest but most informative sets of genes associated with specific clinical features, which then can be interrogated using technologies available in clinical laboratories. Yet another challenge is to determine how well RNA levels of predictive genes correlate with protein levels. A lack of correlation may imply that the predictive property of the gene(s) is independent of gene function. For example, comparisons of messenger RNA and protein levels for the same tumours reported for lung cancer demonstrated that only a small percentage of genes had a statistically significant correlation between the levels of their corresponding proteins and mRNAs²⁷.

Technologies for DNA microarray analysis are still evolving. There is a tendency by manufacturers to favour oligonucleotide- over complementary DNA-based microarrays, and progress has been made on adoption of data analysis standards²⁸. Nevertheless, however perfected DNA microarrays and their analytical tools become for disease profiling, they will not eliminate a pressing need for other types of profiling technologies that go beyond measuring RNA levels, particularly for disease-related investigations. DNA microarrays have limited utility for the analysis of biological fluids and for uncovering assayable biomarkers directly in the fluid. Numerous alterations may occur in proteins that are not reflected in changes at the RNA level, providing a compelling rationale for direct analysis of gene expression at the protein level. As a result, there is substantial interest in developing microarrays or biochips that allow the systematic analysis of thousands of proteins (see review in this issue by Fields and co-workers, page 208).

Unlike DNA microarrays, which provide one measure of gene expression (namely RNA levels), there is a need to implement protein microarray strategies that address the many different features of proteins that can be altered in disease. These include, on the one hand, determination of their levels in biological samples and, on the other, determination of their selective interactions with other biomolecules, such as other proteins, antibodies, drugs or various small ligands. The compelling need for protein chips has led numerous biotechnology companies to devise new strategies for producing biochips that have utility for biomedical investigations. New classes of capture agents include aptamers (SomaLogic, <http://www.somallogic.com/>), ribozymes (Archemix, <http://www.archemix.com/>), partial-molecule imprints (Aspira Biosystems, <http://www.aspirabio.com>) and modified binding proteins (Phylos, <http://www.phylos.com>). For assays of protein interaction, biochips that contain either peptides or proteins are being produced. Peptides may be synthesized in very large numbers directly on the chip²⁹. Alternatively, recombinant proteins may be arrayed and effort is underway to assemble large sets of purified recombinant proteins for microarrays and other applications.

Profiling studies of disease tissue that have used protein microarrays are beginning to emerge. As a model to better understand how patterns of protein expression shape the tissue microenvironment, Knezevic *et al.* analysed protein expression in tissue derived from squamous cell carcinomas of the oral cavity through an antibody

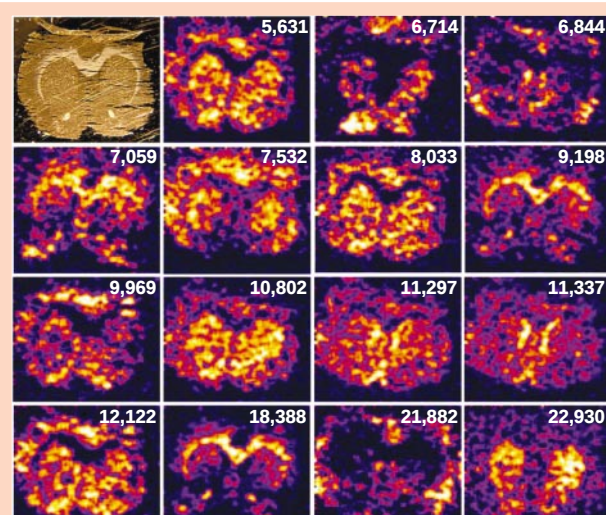


Figure 2 Imaging mass spectrometry. Transverse sections of rat brain were cut, thaw-mounted on the target plate and coated with matrix²¹. A survey scan was performed first with data acquisition taken randomly across the section to generate an average protein profile. Over 200 individual mass peaks were detected in a mass-to-charge (m/z) range of up to 40,000. The figure presents an optical image of the brain section prior to matrix deposition. The section was scanned by acquiring 74×75 points with a resolution of $180 \mu\text{m}$ by averaging spectra produced by 15 laser shots using an automated imaging computer algorithm. In this scan, the intensity of all of the different mass signals was monitored. Fifteen ion-density maps are shown, each obtained for different protein signals; some of these, in particular m/z 6,844, have low intensities. As expected, some proteins were found to be highly specific for a given brain region. This is particularly striking for the density maps of the proteins detected at m/z 5,631 and m/z 18,388, which are almost 'negatives' of each other.

microarray approach for high-throughput proteomic analysis³⁰. Using laser-capture microdissection to procure total protein from specific microscopic cellular populations, they showed that quantitative, and potentially qualitative, differences in expression patterns of multiple proteins within epithelial cells correlated reproducibly with oral-cavity tumour progression. Differential expression of multiple proteins was found in stromal cells surrounding and adjacent to regions of diseased epithelium that correlated directly with tumour progression of the epithelium. Most of the proteins identified in both cell types were involved in signal transduction pathways. Knezevic *et al.* hypothesized therefore that extensive molecular communications involving complex cellular signalling between epithelium and stroma play a key role in driving progression of oral-cavity cancer.

A reverse-phase protein array approach that immobilizes the whole repertoire of a tissue's proteins has been developed³¹. A high degree of sensitivity, precision and linearity was achieved, making it possible to quantify the phosphorylated status of signal proteins in subpopulations of human tissue cells. Using this approach, Paweletz *et al.*³¹ performed a longitudinal analysis of the state of pro-survival checkpoint proteins at the microscopic transition stage from patient-matched, histologically normal prostate epithelium to prostate intraepithelial neoplasia and to invasive prostate cancer. Cancer progression was associated with extensive phosphorylation of the serine/threonine kinase Akt, suppression of apoptosis pathways, and decreased phosphorylation of extracellular signal-regulated kinase (ERK). At the transition from histologically normal epithelium to intraepithelial neoplasia, a statistically significant surge in phosphorylated Akt was observed, together with a concomitant suppression of downstream apoptosis pathways preceding the transition into invasive carcinoma.

A clinically relevant application of protein microarrays is the identification of proteins that induce an antibody response in

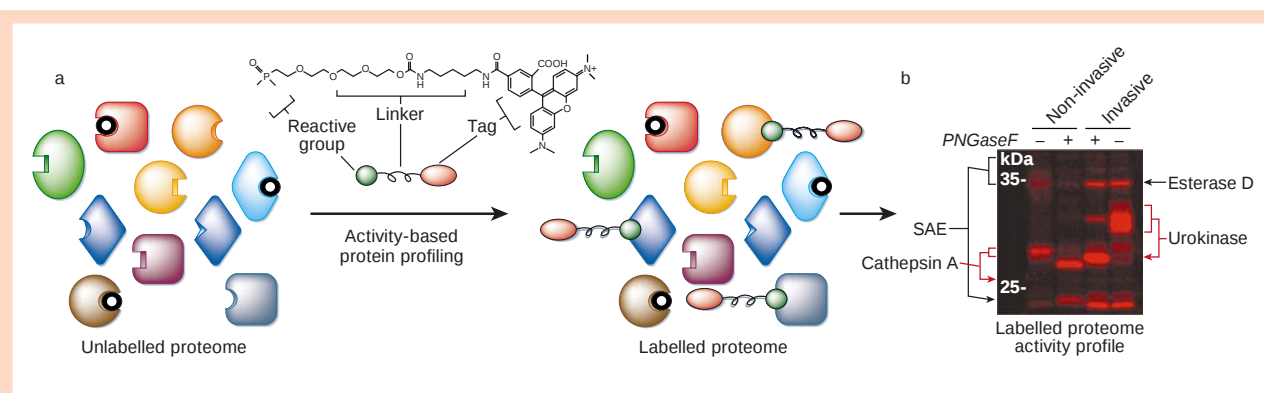


Figure 3 Activity-based protein profiling. This chemical strategy is used to monitor changes in the functional state of enzyme superfamilies directly in complex proteomes⁴⁷. **a**, The method uses chemical probes that comprise three general elements: an active-site-directed reactive group (shown here as an ethoxy fluorophosphonate group that targets the serine hydrolase class of enzymes), a linker (shown here as a polyethylene glycol group) and a tag for the visualization of active enzymes (shown here as a rhodamine group). Active

enzymes are denoted by shapes with open active sites; their inactive counterparts have their active sites shaded in black. **b**, Labelled proteome activity profile. Representative in-gel fluorescence analysis of the secreted proteome (labelled with fluorescent pigment) derived from invasive MUM-2B and non-invasive MUM-2C human melanoma cancer cells. Enzyme activities selectively associated with either invasive cells (for example, urokinase) or non-invasive cells (for example, sialic acid 9-*O*-acetyltransferase or SAE) are noted.

autoimmune disorders³². Microarrays were produced by attaching several hundred proteins and peptides to the surface of derivatized glass slides. Arrays were incubated with patient serum, and fluorescent labels were used to detect autoantibody binding to specific proteins in autoimmune diseases, including systemic lupus erythematosus and rheumatoid arthritis. Such microarrays represent a powerful tool to study immune responses in a variety of diseases, including cancer.

One of the main challenges in making biochips for global analysis of protein expression is the current lack of comprehensive sets of genome-scale capture agents such as antibodies. Another important consideration in protein microarrays is that proteins undergo numerous post-translational modifications that may be crucial to their functions. But these modifications are generally not captured using either recombinant proteins or antibodies that do not distinctly recognize specific forms of a protein. One approach for comprehensive analysis of proteins in their modified forms is to array proteins isolated directly from cells and tissues following protein fractionation schemes³³. Fractions that react with specific probes are within the reach of chromatographic and gel-based separation techniques for resolving their individual protein constituents, and of mass spectrometric techniques for identification of their constituent proteins. Protein microarrays of different types are likely to become commercially available for clinically relevant assays of broad sets of proteins and may well rival DNA microarrays for introduction into the clinical laboratory.

The quest for disease biomarkers using proteomics

There is substantial interest in applying proteomics to the identification of disease markers. Approaches include comparative analysis of protein expression in normal and disease tissues to identify aberrantly expressed proteins that may represent new markers, analysis of secreted proteins in cell lines and primary cultures, and direct serum protein profiling. The potential of mass spectrometry to yield comprehensive profiles of peptides and proteins in biological fluids without the need to first carry out protein separations has attracted interest. In principle, such an approach is highly suited for marker identification because of reduced sample requirements and high throughput.

This approach is currently popularized, particularly for serum analysis, by the technology referred to as surface-enhanced laser desorption/ionization¹. Microlitre quantities of serum from many samples are applied to the surface of a protein-binding plate, with properties to bind a class of proteins. The bound proteins are treated

and analysed by MALDI. The mass spectra patterns obtained for different samples reflect the protein and peptide contents of these samples. Patterns that distinguish between cancer patients and normal subjects with remarkable accuracy have been reported for several types of cancer¹. The main drawbacks of direct analysis of tissues or biological fluids by MALDI are the preferential detection of proteins with a lower molecular mass and the difficulty in determining the identity of proteins owing to post-translational modifications obscuring the correspondence of measured and predicted masses. Occasionally the masses observed match precisely the predicted masses of specific proteins. This was the case in a study of proteins secreted by stimulated CD8 T cells, which led to the identification of the small proteins α -defensin 1, 2 and 3 as contributing to the anti-HIV-1 activity of CD8 antiviral factor³⁴.

A productive approach for the identification of cancer markers has been the analysis of serum for autoantibodies against tumour proteins. There is increasing evidence for an immune response to cancer in humans, demonstrated in part by the identification of autoantibodies against a number of intracellular and surface antigens detectable in sera from patients with different cancer types³⁵. The identification of panels of tumour antigens that elicit an antibody response may have utility in cancer screening, diagnosis or in establishing prognosis, and in immunotherapy against the disease.

There are several approaches for the detection of tumour antigens that induce an immune response³⁵. A number of antigens have been detected by screening expression libraries with patient sera^{36–41} or, more recently, by using a random peptide-library approach⁴². Multiple proteins that induce autoantibodies that are specific for different types of cancer have been identified using 2D gels to separate tumour proteins, followed by western blotting and incubation with patient sera⁴³. For most antigenic proteins identified using this approach, post-translational modifications contributed to the immune response. In a study of lung cancer, sera from 60% of patients with lung adenocarcinoma and from 33% of patients with squamous-cell lung carcinoma, but from none of the non-cancer controls, exhibited immunoglobulin- γ -based reactivity against proteins identified as glycosylated annexins I and II⁴⁴. Microarrays that contain proteins derived from tumour cells have the potential of substantially accelerating the pace of discovery of tumour antigens and yielding a molecular signature for immune responses directed against protein targets in different types of cancer³³.

The increased emphasis on proteomics for disease investigations is stimulating a reassessment of strategies for sample procurement and preservation to render them compatible with proteomics,

because of the inherent instability of proteins. For example, the manner in which biological fluids such as serum or plasma are collected is not ideally suited for proteomics. There is a need to reduce protein degradation and other forms of modifications that may substantially alter protein content and interfere with global profiling.

Disease-related functional proteomics

Although data obtained by various expression proteomics strategies have functional relevance by uncovering altered levels or post-translational modification states of proteins in disease, additional technologies are needed for more direct functional analysis. This is exemplified by the need to analyse protein complexes and their disruption in disease, to assay in a high-throughput fashion the activity of various classes of proteins, and to manipulate the levels and activities of individual proteins, in a cellular context, to determine their role in different biological processes and disease states.

Various strategies are currently in use for studies of protein complexes and protein–protein interactions (see review in this issue by Fields and co-workers, page 208). So far, such strategies have been applied to disease investigations on a limited basis, with a relatively narrow focus on particular complexes. For example, using affinity pull-down assays, 2D gels and mass spectrometry, myocardial protein kinase C ϵ (PKC ϵ) was found to be associated physically with at least 36 other proteins with a multitude of functions. Cardioprotection induced by activation of PKC ϵ was found to be coupled with dynamic modulation and recruitment of PKC ϵ -associated proteins⁴⁵. Previously unrecognized functions of PKC ϵ have relevance to heart function and heart disease. Clearly, at the present time, most systematic studies of protein–protein interactions have dealt with normal or physiological states, as the field is still in its early stages and the merits of various technologies, particularly for disease investigations, have yet to be fully appreciated.

Likewise, there has been limited application of activity-based proteomics to disease investigations, although the field itself is still in its early stages^{46–48}. Activity-based assays of normal and disease tissue have substantial relevance to the study of disease, but probably will be limited to the analysis of one class of proteins at a time. The potential contribution of this technique to the measurement of global dynamics in protein function is illustrated by the application of a chemical proteomics strategy to quantitatively compare enzyme activities across normal and disease tissue⁴⁷ (Fig. 3). A global analysis of the activity, subcellular distribution and glycosylation state for the serine hydrolase superfamily in a panel of human breast and melanoma cell lines resulted in the identification of a cluster of proteases lipases and esterases that distinguished cancer lines based on tissue of origin. Remarkably, the majority of these enzyme activities were downregulated in the most invasive cancer lines examined, which instead upregulated a distinct set of secreted and membrane-associated enzyme activities.

Approaches are being developed or implemented to allow direct protein manipulations other than by decreasing or increasing overall amounts of protein, as can be done by gene manipulations. One advantage of these approaches is the ability to inactivate a specific site in a protein in a time-dependent and localization-restricted manner, as with chromophore-assisted laser inactivation (CALI). Such inactivation allows assessment of the disease relevance of the site and provides a means to revert a cellular phenotype to a more normal state⁴⁹ (Fig. 4). Inactivation of protein targets involved in different signal-transduction pathways in cancer using CALI has been demonstrated⁴⁹.

Contributions of proteomics to studies of pathogens

Despite earlier predictions to the contrary, infectious diseases remain as a leading cause of death worldwide. A complicating factor in therapy for infectious disease is the development of resistance to commonly used drugs (for example, as has occurred in tuberculosis), which heightens the need for developing effective new therapies.

Table 1 Targeted proteomic approaches to microbial pathogens

- Characterization of submicrobial proteomes (for example, secreted proteins, surface proteins and immunogenic proteins)
- Comparative analysis of different strains
- Comparative analysis of different physiological states
- Identification of proteins related to pathogenicity
- Identification of proteins involved in host–pathogen interactions
- Evaluation of mechanisms of action of antimicrobials

Interest in the application of proteomics to microbiology goes back at least two decades, with the pioneering work of Fred Neidhardt to characterize protein expression patterns in *Escherichia coli* under different growth conditions⁵⁰. The complete sequencing of a number of microbial genomes has provided a framework for identifying proteins encoded in these genomes using mass spectrometry. A case in point is the sequencing of the genome of the malaria parasite *Plasmodium falciparum*, which has provided a basis for conducting comparative proteomics studies of this pathogen, leading to the identification of new potential drug and vaccine targets^{51,52}. Aside from comprehensive identification of microbial proteins, proteomics is relevant to numerous aspects of microbial disease pathogenesis and treatment^{53–57} (Table 1).

Contribution of proteomics to drug development

There is currently a burgeoning interest in proteomics on the part of the pharmaceutical industry, evidenced by implementation of proteomics programmes by most major pharmaceutical companies. The notion has been advanced that, as the vast majority of drugs target proteins, proteomics should have substantial utility for drug development. But the industry has so far adopted a cautious attitude, and it is too early to make a critical assessment of the contributions of proteomics to drug development, relative to other approaches. The caution stems from the prior heavy investment in genomics and other approaches and some uncertainty surrounding the adequacy and scalability of proteomics to meet the needs of the pharmaceutical industry. Provided suitable technology platforms become available, the use of proteomics may permeate numerous aspects of drug development, by identifying new targets and facilitating assessment of drug action and toxicity both in the preclinical and clinical phases.

Several published studies illustrate the application of functional proteomics for identification of regulated targets in specific pathways. Lewis *et al.*⁵⁸ have combined functional proteomics with selective activation and inhibition of mitogen-activated protein kinase (MAPK) kinase (MKK), in order to identify cellular targets regulated by the MKK/ERK cascade. Twenty-five targets of this signalling pathway were identified, of which only five were previously characterized as MKK/ERK effectors. The remaining targets suggest new roles for this signalling cascade in cellular processes of nuclear transport, nucleotide excision repair, nucleosome assembly, membrane transport and cytoskeletal regulation.

In another study to identify proteases most suitable for drug targeting, an automated microtitre-plate assay was modified to allow detection of the four main classes of proteases in tissue samples (matrix metalloproteases, cathepsins, and the cell serine proteases, trypsin and chymase)⁵⁹. Fifteen sets of colorectal carcinoma biopsies representing primary tumour, adjacent normal colon and liver metastases were screened for protease activity. Matrix metalloproteases were expressed at higher levels in the primary tumour than in adjacent normal tissue. The mast cell proteases, in contrast, were found at very high levels in adjacent normal tissue, but were not detectable in the metastases. Cathepsin B activity was significantly higher in the primary tumour, and highest in the metastases. The proteases detected by activity assays were then localized in biopsy sections by immunohistochemistry. Mast cell proteases were abundant in adjacent normal tissue, because of infiltration of the lamina propria by mast cells. Matrix metalloproteases were localized to the

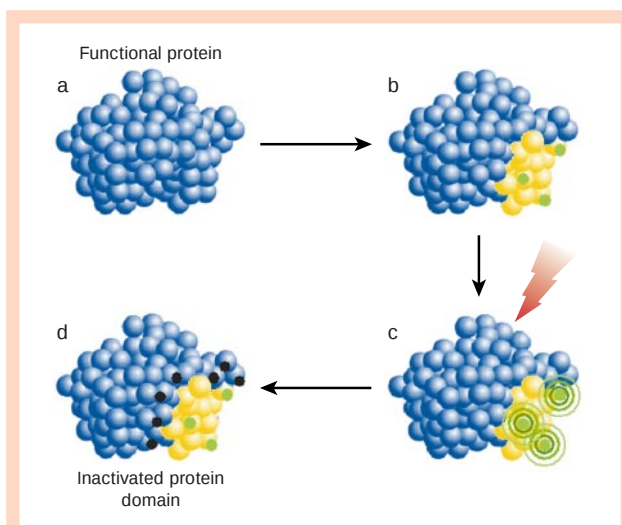


Figure 4 Chromophore-assisted laser inactivation. This technology involves the generation of short-lived radicals that induce covalent modifications at spatially restricted sites on a protein⁴⁹. A transient functional inactivation occurs if the radicals modify amino acids of the target protein that have a functional role. A target protein (a) is complexed with a ligand (yellow; for example, antibody, antibody fragments such as scFv and Fab, peptide, nucleic acid aptamers or small molecules) labelled with dye molecules (green; for example, malachite green or fluorescein) (b). The complex is irradiated with laser or incoherent light (red lightning), leading to the generation of reactive species (concentric circles) that travel a short distance (c). The reactive species in turn lead to modifications (black) of nearby amino acids (d). If the modified amino acids are responsible for a function, that particular functional protein domain will be inactivated, leaving the other functional domains intact.

tumour cells themselves, whereas cathepsin B was expressed predominantly by macrophages at the leading edge of invading tumours.

Such activity-based screening provides a basis for selecting targets in the development of inhibitors to specific proteases. Protein biochips potentially could provide a high-throughput platform for target identification. Biochips developed for interaction studies could be important in lead-compound optimization and could accelerate drug development by allowing efficient evaluation of lead compounds for specificity and selectivity in binding to drug targets. Proteomics also may provide increased efficiency of clinical trials through the availability of biologically relevant markers for drug efficacy and safety.

Organizing proteomics initiatives

It is clear that while some progress has been made in disease proteomics, the field is still in its infancy. Knowledge of the sequence of the human genome has provided a framework for genomic approaches to unravel disease processes. A similar knowledge of the human proteome is currently lacking. Developing a comprehensive knowledge framework of the proteome is considerably more complex than sequencing the human genome. Ideally, such a proteome framework would encompass knowledge of all human proteins, from their sequence to their post-translational modifications, to their interactions among each other, their cellular and sub-cellular distribution, and their temporal pattern of expression. Although such an exhaustive framework will not materialize in the foreseeable future, more modest goals may well be within reach.

To that effect, there is a need to begin an organized effort, the goals of which include developing an infrastructure in proteomics that would substantially facilitate unravelling the complexity of the proteome in health and in disease. The Human Proteome Organisation (HUPO, <http://www.hupo.org>) was founded to regroup scientists in the public and private sectors engaged throughout the world in

various aspects of proteomics. HUPO's mission is threefold: to consolidate national and regional proteome organizations into a worldwide organization; to engage in scientific and educational activities to encourage the spread of proteomics technologies and disseminate knowledge pertaining to the human proteome and that of model organisms; and to assist in the coordination of public proteome initiatives aimed at characterizing specific tissue and cell proteomes. Initiatives currently in the pilot phase include an international effort to identify proteins detectable in normal serum and plasma and their range of variation with age, ethnicity and physiological state, and a liver proteome study to identify proteins expressed in the liver. These initiatives have attracted substantial interest and will be integrated with efforts in protein informatics to achieve data standardization on the one hand, and data curation on the other.

Concluding remarks

Proteome alterations in disease may occur in many different ways that are not predictable from genomic analysis, and it is clear that a better understanding of these alterations will have a substantial impact in medicine. A useful repertoire of proteomics technologies is currently available for disease-related applications, although further technological innovations would be beneficial to increase sensitivity, reduce sample requirement, increase throughput and more effectively uncover various types of protein alterations such as post-translational modifications. The use of these technologies will likely expand substantially, particularly to meet the need for better diagnostics and to shorten the path for developing effective therapy. □

doi:10.1038/nature01514

- Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C. & Liotta, L. A. Clinical proteomics: translating benchside promise into bedside reality. *Nature Rev. Drug Discov.* **1**, 683–695 (2002).
- Hanash, S. 2-D or not 2-D—is there a future for 2-D gels in proteomics? Insights from York proteomic meeting. *Proteomics* **1**, 635–637 (2001).
- Hanash, S. M., Madoz-Gurpide, J. & Misk, D. E. Identification of novel targets for cancer therapy using expression proteomics. *Leukemia* **16**, 478–485 (2002).
- Van Eyk, J. E. Proteomics: unraveling the complexity of heart disease and striving to change cardiology. *Curr. Opin. Mol. Therapeut.* **3**, 546–553 (2001).
- Li, X. P. *et al.* A two-dimensional gel electrophoresis database of rat heart protein. *Electrophoresis* **20**, 891–897 (1999).
- Evans, G., Wheeler, C. H., Corbett, J. M. & Dunn, M. J. Construction of HSC-2D PAGE: a two-dimensional gel electrophoresis database of heart proteins. *Electrophoresis* **18**, 471–479 (1997).
- van Der Velden, J. *et al.* Effects of calcium, inorganic, phosphate, and pH on isometric force in single skinned cardiomyocytes from donor and failing human hearts. *Circulation* **104**, 1140–1146 (2001).
- Arrell, D. K., Neverova, I., Fraser, H., Marbán, E. & Van Eyk, J. E. Proteomic analysis of pharmacologically preconditioned cardiomyocytes reveals novel phosphorylation of myosin light chain 1. *Circ. Res.* **89**, 480–487 (2001).
- Ping, P., Zhang, J., Pierce, W. M. & Bolli, R. Functional proteomic analysis of protein kinase C ϵ signaling complexes in the normal heart and during cardioprotection. *Circ. Res.* **88**, 59–62 (2001).
- Heinke, M. Y. *et al.* Protein changes observed in pacing-induced heart failure using two-dimensional electrophoresis. *Electrophoresis* **19**, 2021–2030 (1998).
- Westbrook, J. A., Yan, J. X., Wait, R., Welson, S. Y. & Dunn, M. J. Zooming-in on the proteome: very narrow-range immobilized pH gradients reveal more protein species and isoforms. *Electrophoresis* **22**, 2865–2871 (2001).
- Hoving, S. *et al.* Preparative two-dimensional gel electrophoresis at alkaline pH using narrow range immobilized pH gradients. *Proteomics* **2**, 127–134 (2002).
- Langen, H. *et al.* Two-dimensional map of the proteome of *Haemophilus influenzae*. *Electrophoresis* **21**, 411–429 (2000).
- Zuo, X. & Speicher, D. W. Comprehensive analysis of complex proteomes using microscale solution isoelectrofocusing prior to narrow pH range two-dimensional electrophoresis. *Proteomics* **2**, 58–68 (2002).
- Patton, W. F. Detection technologies in proteome analysis. *J. Chromatogr. B* **771**, 3–31 (2002).
- Zhou, G. *et al.* 2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers. *Mol. Cell. Proteomics* **1**, 117–124 (2001).
- Gagnon, E. *et al.* Endoplasmic reticulum-mediated phagocytosis is a mechanism of entry into macrophages. *Cell* **110**, 119–131 (2002).
- Sabarth, N. *et al.* Identification of surface proteins of *Helicobacter pylori* by selective biotinylation, affinity purification, and two-dimensional gel electrophoresis. *J. Biol. Chem.* **277**, 27896–27902 (2002).
- Shin, B. K. *et al.* Global profiling of the cell surface proteome of cancer cells uncovers an abundance of proteins with chaperone function. *J. Biol. Chem.* DOI: 10.1074/jbc.M210455200 (2002).
- Brivio, M. *et al.* Integrated microfluidic system enabling (bio)chemical reactions with on-line MALDI-TOF mass spectrometry. *Anal. Chem.* **74**, 3972–3976 (2002).
- Stoeckli, M., Chaurand, P., Hallahan, D. E. & Caprioli, R. M. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature Med.* **7**, 493–496 (2001).

22. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
23. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
24. Brenton, J. D., Aparicio, S. A. & Caldas, C. Molecular profiling of breast cancer: portraits but not physiognomy. *Breast Cancer Res.* **3**, 77–80 (2001).
25. Gruvberger, S. *et al.* Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **61**, 5979–5984 (2001).
26. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
27. Chen, G. *et al.* Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors. *Clin. Cancer Res.* **8**, 2290–2305 (2002).
28. Opinion. Microarray standards at last. *Nature* **419**, 323 (2002).
29. Pellois, J. P. *et al.* Individually addressable parallel peptide synthesis on microchips. *Nature Biotechnol.* **20**, 922–926 (2002).
30. Knezevic, V. *et al.* Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics* **1**, 1271–1278 (2001).
31. Pawelcz, C. P. *et al.* Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**, 1981–1989 (2001).
32. Robinson, W. H. *et al.* Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature Med.* **8**, 295–301 (2002).
33. Madoz-Gurpide, J., Wang, H., Misek, D. E., Brichory, F. & Hanash, S. M. Protein based microarrays: a tool for probing the proteome of cancer cells and tissues. *Proteomics* **1**, 1279–1287 (2001).
34. Zhang, L. *et al.* Contribution of human α -defensin 1, 2 and 3 to the anti-HIV-1 activity of CD8 antiviral factor. *Science* **298**, 995–1000 (2002).
35. Hanash, S. Harnessing immunity for cancer marker discovery. *Nature Biotechnol.* **21**, 37–38 (2003).
36. Stockert, E. *et al.* A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. *J. Exp. Med.* **187**, 1349–1354 (1998).
37. Gourevitch, M. M. *et al.* Polymorphic epithelial mucin (MUC-1)-containing circulating immune complexes in carcinoma patients. *Br. J. Cancer* **72**, 934–938 (1995).
38. Gure, A. O. *et al.* Human lung cancer antigens recognized by autologous antibodies: definition of a novel cDNA derived from the tumor suppressor gene locus on chromosome 3p21.3. *Cancer Res.* **58**, 1034–1341 (1998).
39. Yamamoto, A., Shimizu, E., Ogura, T. & Sone, S. Detection of auto-antibodies against L-myc oncogene products in sera from lung cancer patients. *Int. J. Cancer* **22**, 283–289 (1996).
40. Soussi, T. The humoral response to the tumor-suppressor gene product p53 in human cancer: implications for diagnosis and therapy. *Immunol. Today* **17**, 354–356 (1996).
41. Old, L. J. & Chen, Y. T. New paths in human cancer serology. *J. Exp. Med.* **187**, 1163–1167 (1998).
42. Mintz, P. J. *et al.* Fingerprinting the circulating repertoire cancer patients. *Nature Biotechnol.* **21**, 57–63 (2003).
43. Le Naour, F. Contribution of proteomics to tumor immunology. *Proteomics* **1**, 1295–1302 (2001).
44. Brichory, F. M. *et al.* An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. *Proc. Natl Acad. Sci. USA* **98**, 9824–9829 (2001).
45. Vondriska, T. M. & Ping, P. Functional proteomics to study protection of the ischaemic myocardium. *Expert Opin. Therapeut. Targets* **6**, 563–570 (2002).
46. Borodovsky, A. *et al.* Chemistry-based functional proteomics reveals novel members of the deubiquitinating enzyme family. *Chem. Biol.* **9**, 1149–1159 (2002).
47. Jessani, N., Liu, Y., Humphrey, M. & Cravatt, B. F. Enzyme activity profiles of the secreted and membrane proteome that depict cancer cell invasiveness. *Proc. Natl Acad. Sci. USA* **99**, 10335–10340 (2002).
48. Greenbaum, D. *et al.* Chemical approaches for functionally probing the proteome. *Mol. Cell. Proteomics* **1**, 60–68 (2002).
49. Rubenwolf, S. *et al.* Functional proteomics using chromophore-assisted laser inactivation. *Proteomics* **2**, 241–246 (2002).
50. VanBogelen, R. A., Schiller, E. E., Thomas, R. D. & Neidhardt, F. C. Diagnosis of cellular states of microbial organisms using proteomics. *Electrophoresis* **20**, 2149–2159 (1999).
51. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537–542 (2002).
52. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).
53. Nilsson, C. L. Bacterial proteomics and vaccine development. *Am. J. Pharmacogenomics* **2**, 59–65 (2002).
54. Eymann, C., Homuth, G., Scharf, C. & Hecker, M. *Bacillus subtilis* functional genomics: global characterization of the stringent response by proteome and transcriptome analysis. *J. Bacteriol.* **184**, 2500–2520 (2002).
55. Haas, G. *et al.* Immunoproteomics of *Helicobacter pylori* infection and relation to gastric disease. *Proteomics* **2**, 313–324 (2002).
56. Reid, S. D. *et al.* Postgenomic analysis of four novel antigens of group A streptococcus: growth phase-dependent gene transcription and human serologic response. *J. Bacteriol.* **184**, 6316–6324 (2002).
57. Antelmann, H., Yamamoto, H., Sekiguchi, J. & Hecker, M. Stabilization of cell wall proteins in *Bacillus subtilis*: a proteomic approach. *Proteomics* **2**, 591–602 (2002).
58. Lewis, T. S. *et al.* Identification of novel MAP kinase pathway signaling targets by functional proteomics and mass spectrometry. *Mol. Cell* **6**, 1343–1354 (2000).
59. McKerrow, J. H. *et al.* A functional proteomics screen of proteases in colorectal carcinoma. *Mol. Med.* **6**, 450–460 (2000).

Acknowledgements I thank P. Chaurand, R. Caprioli, G. Omenn, B. Cravatt, N. Jessani, R. Kuick, L. Ilag, E. Gulari for their insightful comments, stimulating discussions and other contributions to the preparation of this review.

Biomedical informatics for proteomics

Mark S. Boguski* & Martin W. McIntosh†

*Human Biology Division, and †Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, PO Box 19024, Seattle, Washington 98109, USA (e-mail: mboguski@fhcrc.org; mmcintos@fhcrc.org)

Success in proteomics depends upon careful study design and high-quality biological samples. Advanced information technologies, and also an ability to use existing knowledge to the full, will be crucial in making sense of the data. Despite its genome-scale potential, proteome analysis is at a much earlier stage of development than genomics and gene expression (microarray) studies. Fundamental issues involving biological variability, pre-analytic factors and analytical reproducibility remain to be resolved. Consequently, the analysis of proteomics data is currently informal and relies heavily on expert opinion. Databases and software tools developed for the analysis of molecular sequences and microarrays are helpful, but are limited owing to the unique attributes of proteomics data and differing research goals.

The subtitle of a recent conference on the Human Proteome Project asserted that “Genes Were Easy”¹. Depending upon one’s perspective, this statement might elicit feelings of hubris, envy or fear about the challenges and complexities of another ostensible paradigm shift in biomedical research. We have transitioned rapidly from the momentary comfort of a large, but finite and complete human genome to a seemingly infinite biological universe of post-transcriptional complexities^{2–4}.

Proteomics is often referred to as a ‘post-genome’ science, but its antecedents actually predate the Human Genome Project by two to three decades and developed along different intellectual lines^{5,6}. Bioinformatics, although enjoying its ascendancy during the earlier days of genome sequencing^{7,8}, also traces its roots to a time long before the development of cloning and sequencing technologies, when protein primary structures were determined experimentally and not derived routinely and automatically from conceptual translations of coding DNA^{9–11}. Although medical informatics¹² has until recently been largely detached from bioinformatics, the emergence of clinical genomics and proteomics increasingly requires the integrated analysis of genetic, cellular, molecular and clinical information and the expertise of pathologists, epidemiologists and biostatisticians.

Proteomics is the latest functional genomics¹³ technology to capture our imagination and it is instructive to review some lessons learned during the earlier adoption of another functional genomics technology, namely gene expression analysis using microarrays and similar technologies^{14,15}. Study design and sample quality, databases, data analysis and data standards are discussed with special emphasis on human plasma and serum proteomics (Box 1) because of the enormous potential of these studies to advance clinical diagnostics and therapeutic monitoring (see review in this issue by Hanash, page 226).

There are many implications of biomedical informatics for proteomics, including multiple platform technologies (for example, two-dimensional polyacrylamide gel electrophoresis, mass spectrometry, protein and antibody arrays), laboratory information-management systems, medical records systems, and documentation of clinical trial

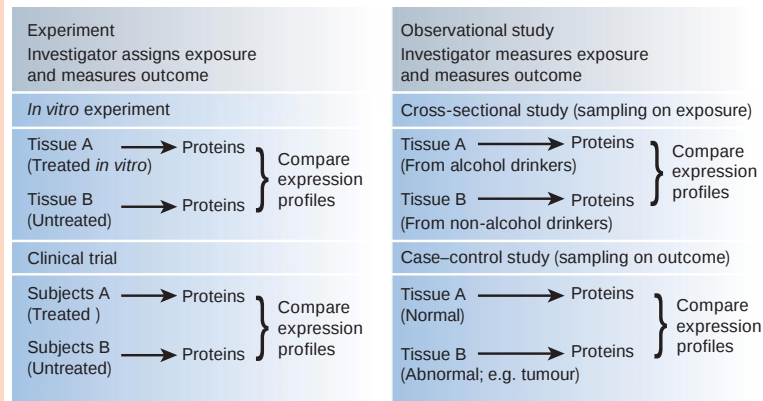
results for regulatory agencies. In the present work, we confine our discussions to mass spectrometry-based proteomics (and see accompanying review by Aebersold and Mann, page 198), and to study design and data resources, tools and analysis in a research setting.

Study design and sample quality

Potter¹⁶ describes four study designs illustrating, for example, the difference between a clinical trial (experiment) and cross-sectional (observational) study (Fig. 1). There are critical differences between experimental biomedicine and epidemiological studies and in many microarray gene expression studies “the distinction between observational and experimental designs is not made”¹⁶. The same arguments hold true for proteomics studies. Indeed, most gene expression and proteomic analyses involving human specimens will, of necessity, be observational studies and this fact immediately raises the key issues of possible biases and confounding factors in the populations from which the samples are drawn. Plasma and serum proteomics, defined as the discovery and utilization of biomarkers in clinical blood specimens, provides an illustrative case in point.

Human plasma and serum proteomics may be particularly susceptible to observational biases because any confounding factor (such as smoking, diet or ascertainment bias) could conceivably cause a phenotypic response that might be confused with a specific characteristic of the disease process under study. Without careful sample ascertainment and/or the availability of detailed sample annotation, the conclusion of any such study can be misleading. For example, consider the task of acquiring serum samples in an attempt to identify, through proteomic analysis, diagnostic biomarkers that can differentiate cancer patients from healthy subjects. It is common and convenient to collect the disease specimens during surgery, whereas the control subjects do not typically donate their specimens in the operating room. Such a design results in complete confounding between specimen ascertainment and disease status and so it is impossible to determine whether any finding reflects a marker for disease or instead is a marker for pre-operative fasting, anaesthesia, psychological stress, or some other uncontrollable confounding phenomenon. Indeed, in quantitative terms, it is even known that systematic

Figure 1 Experimental versus observational study. These study types differ by the manner in which an exposure, or treatment, is assigned to the study subjects. Assignment in experimental studies is under the control of the experimenter, whereas they have no control over treatment assignment in observational studies. Many clinical genomics studies are retrospective and observational, relying upon data from patient medical records to provide information on the relevant phenotype as well as pre-analytical variables and potential confounders. (Figure reproduced from ref. 16.)



differences in tourniquet application time, exercise, and whether the sample is obtained while the subject is sitting or recumbent can each individually induce a change in total protein concentrations by $\pm 10\%$ (ref. 17).

It is at least as important for proteomics researchers to understand basic epidemiology as it is to understand complex analytic algorithms. Although pre-analytical errors can conceivably occur in samples of any biological nature or origin, here we have highlighted human serum proteome analysis because the nature of protein discovery work in sera means that confounding variables that affect any tissue or pathway may complicate study findings. Moreover, unlike traditional studies with low-dimensional measurements, none of the analytic methods we found allow any possibility to adjust for confounding even if confounders are annotated. Presently, careful design and specimen ascertainment may be the only way to have confidence in study findings with human subjects.

Two additional issues, associated with specimens of almost any kind, are sample quality and number. Quality involves both the preservation of molecular features (such as intact and representative messenger RNAs and proteins) and the assurance of both inter- and intra-sample homogeneity. For example, Huang *et al.*¹⁸ have shown that the duration of ischaemia associated with surgical resection of tissues has significant effects on gene expression. Craven and Bank^{19,20} describe some aspects of sample heterogeneity in proteomics, and methods (such as laser-capture microdissection) to address it.

In referring to functional genomics technologies and their relevance in clinical medicine, Margolin has admonished²¹ that “Scientists...need to avoid the tendency, often driven by the high price of some of the newer techniques, of running under-controlled experiments or experiments with fewer repeated conditions than would have been accepted with standard techniques.” The same caveat applies for proteomics research, but perhaps even more so because a framework to estimate efficient sample sizes has yet to be determined, and the nature of the technology creates substantial challenges to progress in achieving this goal. For example, with microarrays, because the number of interrogations is determined pre-experimentally by the number of genes or gene-specific probes on the array, the confidence in declaring one or a group of genes as differentially expressed can be quantified (using statistical *P* values) by reporting or controlling the rate of false identification²². Proteomics discovery with complex mixtures like sera has no such *a priori* enumeration of targets, and the discovery procedure is iterative and far more informal. The lack of a described procedural structure at this time makes it difficult to make any statement about the confidence of any finding.

Protein databases

Collections of protein sequences date back to the 1960s²³, preceding GenBank by nearly 20 years²⁴. Since the early 1990s, important

utilitarian goals of protein databases have included minimal redundancy, maximal annotation and integration with other databases²⁵. These principles continue to be stressed today²⁶. For both historical and practical reasons, current molecular sequence databases are designed to represent a comprehensive ‘parts list’ of an organism’s genome, that is, the genes and all of the proteins they encode, and protein ‘families’ are usually classified according to their evolutionary history inferred from sequence homology. These databases are thus excellent tools for gene discovery, comparative genomics and molecular evolution, but there is much work to be done to even minimally serve the needs of proteomics and integrative biological science^{27,28}.

Today’s principal protein databases emphasize molecular and cellular features and annotation and are not well suited to represent physiology. For example, there are approximately 500 known human serum proteins^{17,29} with extensive information about normal and abnormal ‘reference’ values in health and disease³⁰. But simple searches of the popular protein databases, SWISSPROT and LocusLink²⁶, using the terms ‘human serum’ and ‘serum’ yielded only 36–44 and 81–268 matches, respectively, and among the latter are many false positives. Similarly, 68–84 and 457–3,850 proteins are retrieved when ‘human plasma’ and ‘plasma’ are used to search SWISSPROT and LocusLink, respectively, again with many false positives.

This is just one example of the fact that there is no reliable or satisfying way to retrieve groups of proteins based upon well-known pathways or functional classifications (for example, coagulation, complement fixation or proteinase inhibitors). Furthermore, annotations about post-translational modifications are sparse and difficult to locate in any consistent way, although some progress is being made³¹. There is also the challenge of distinguishing annotations based upon modifications predicted from protein motifs compared with those based upon direct experimental evidence.

A more ideal database for plasma proteome studies would classify proteins from a functional, rather than an evolutionary, viewpoint (perhaps based upon an updated version of the Putnam classification, as discussed in ref. 32). Such a database would also annotate protein concentrations (and other practically measurable attributes) compared with normal ranges of values in reference samples. Attention to emerging data standards (Box 2) will also be important.

Protein identification by database searching

Until recently, the overarching purpose of database similarity searching was the sensitive detection of sequence homologues, regardless of the species or remoteness of the relationship, in order to infer similarity of function from similarity of sequence and/or to study the evolution of protein families or domains. The specific aims of most proteomics studies are different and therefore require different strategies and tools. For example, in the analysis of human serum,

Box 1

Glossary

Case-control and cohort study. These observational studies differ in the way study subjects are selected. Case-control studies select study subjects based on presence (cases) or absence (controls) of the phenotype (for example, disease) of interest. Cohort studies select participants based on the presence or absence of a risk factor of interest and subjects are followed over time for the development of an outcome of interest.

Confounder/Confounding. A confounder is a variable that distorts an apparent relationship between an exposure and a phenotype of interest. Confounding occurs when the relationship between an outcome (for example, disease) and an exposure of interest cannot be distinguished from other variables that also correlate with the outcome.

Plasma and serum. Plasma is the fluid, non-cellular portion of blood; serum is the protein solution remaining after blood or plasma has been allowed to coagulate. Serum thus lacks clotting-factor proteins. Blood samples are often treated with preservatives, anticoagulants and other additives prior to transport and storage or processing. These are some of the pre-analytical variables that may affect subsequent analyses.

Pre-analytical variables. These refer to those factors, both known and unknown, that may be present in a subject or may arise in any of the steps prior to a laboratory test and data analysis. Examples include genotype, physiological attributes such as age, gender, reproductive status, lifestyle effects (for example, diet or smoking), drugs and specimen collection, handling and processing protocols. Uncontrollable variables must be well understood in order to be able to separate their effects from the object or process under study. Most errors in clinical laboratory tests are known to occur in the pre-analytical phase^{55,56}.

Randomized clinical trial. An experimental study in which treatments are randomly assigned to subjects as a method to prevent treatment choice from being confounded.

one is interested in identifying proteins that are not normally present and/or variances in the concentrations of the normal constituents. The object of a database search in this case is to find an exact, or nearly exact, match between subsequences (peptide fragments) of serum components and those proteins encoded by the large (but finite) human genome. Weak similarities and interspecies matches are not pertinent, except in the case of 'foreign' proteins encoded by infectious organisms and parasites that may be released into the circulation. Statistical significance is important, but not in the sense of the probability that two sequences are related by chance. Rather, one is seeking an answer to the question of whether the presence or absence of a particular protein, at a particular concentration, deviates significantly from a normal range of values. If the condition is met, one is then interested in attempting to demonstrate a significant correlation between this protein and a risk factor or outcome of interest. One must bear these purposes in mind when attempting to use existing databases and search tools in a proteomics context.

There are several approaches that utilize mass spectrometry for protein and peptide analysis^{33,34}. These include analytic peptide-mass fingerprinting, *de novo* sequence interpretation and comparative analysis of actual spectra with predicted spectra of peptide sequences from a protein database. The first and third of these methods use comparisons against a database and the reliability of any database search depends on the accuracy and resolution of data, quality of the sequence database, and that of the scoring algorithm used. The accuracy of the input data is affected by many factors that are unique to mass spectrometry compared with DNA sequencing and conceptual translations to protein^{33,34}. For example, co- and post-translational

modifications^{2,35} of amino acid residues obviously affect the masses of real peptides and cannot be predicted consistently or reliably for virtual peptide sequences, although some search engines use error-tolerant heuristics in an attempt to take potential modifications into account^{36,37}. Additionally, the effects of inaccuracy and discontinuity in both expressed sequence tag data³⁸ and genomic data³⁹, and thus in their encoded peptide sequences, have received some attention.

Selecting which of many candidate spectra is correct involves scoring the similarity of the observed and predicted spectra. Detailed consideration of specific scoring algorithms is beyond the scope of this review and more specific descriptions are found elsewhere^{40–44}. In general, each scoring algorithm designates a quantity related to the probability that the candidate peptide could have produced the observed spectrum by chance. When the number of peptides to identify is small it is feasible for a skilled operator to evaluate all high-scoring candidate peptides manually and make assignments using their expert opinion. For each possible peptide spectrum this score is commonly used to rank the candidate peptides. But manual scoring for complex mixtures is not feasible. Instead, it is common to use the scoring algorithm to rank the candidates and assign only the highest scoring of all. This of course makes automated proteomics highly dependent on the quality of the scoring algorithm used.

Moreover, automated identification based on ranking peptides by their scores is not directly analogous to the well established procedure of ranking expressed genes on microarrays based on their *P* values^{45,46}, because peptide scores are not true *P* values even though they may fall between 0 and 1. Consider a simple example of three candidate peptides labelled PA, PB and PC, which together produce any of the three possible spectra S1, S2 and S3. We consider PB intermediate to PA and PC in that it may produce any of the three spectra (each with a probability of 1/3), but PA may produce either S1 or S2 (each with a probability of 1/2) but not S3 (probability of 0). Likewise, PC may produce either S2 or S3 (each with a probability of 1/2) but not S1 (probability of 0). Now with each possible experimental observation — S1, S2 or S3 — consider which peptide will achieve the highest-ranking score. The score for PB can never achieve a value higher than 1/3 because each of the three spectra are equally likely and the values of PA and PB will score either 0 or 1/2. Indeed, because all observations of at least one PA or PC will score 1/2, peptide PB will never achieve the highest rank (Table 1). Thus, even when considering a mixture rich in peptide PB, the latter will never achieve the highest rank but will instead be misidentified as PA or PC.

Automated peptide identification can be improved if the number of peptide choices (that is, the complexity of the mixture) is reduced. For instance, in the example above, peptide PB could possibly achieve the highest rank if either PA or PC is eliminated as a possible alternative. In simple mixtures, human operators can reduce the complexity by auditing the highest-ranking peptides and using their informal expert opinion to eliminate some of the highest-ranking peptides. Because manual review such as this is not feasible with highly complex mixtures such as sera, some investigators have begun to develop methods to formalize expert opinion and use it in more complex scoring algorithms that can automatically eliminate, or reduce in rank, peptides that would otherwise achieve a high rank. For example, Bafna *et al.*⁴⁰ give an example of how experienced spectrometrists, recognizing in the spectrum so-called neutral losses of water or ammonia from side chains of amino acids, can distinguish among peptide candidates that possess similar high

Table 1 Scoring and peptide ranking in a simple mixture of three peptides

Peptide	Possible spectra	Peptide score (probability) for observed spectra		
		S1	S2	S3
PA	S1, S2	1/2	1/2	0
PB	S1, S2, S3	1/3	1/3	1/3
PC	S2, S3	0	1/2	1/2
Peptide assignment		PA	PA or PC	PC

Box 2

Data standards

There are numerous examples in information management and processing where the existence of multiple and/or specialized file formats has hindered accessibility, information exchange and integration. The functional genomics (microarray) field provides a pertinent model for the development of standards that greatly enhance the opportunities for data access and exchange, data integration and meta-analysis^{57,58}. Adherence to the MIAME standard (for 'minimum information about a microarray experiment') for microarray data is now required for manuscript submission to all *Nature* journals⁵⁹ and *Science* also supports this "evolving standardization"⁶⁰. The Human Proteome Organisation is currently engaged in a proteomics standards initiative⁶¹ to develop formats for mass spectrometry and protein-protein interaction data and annotation. These formats use eXtensible Markup Language (XML), which is an Internet standard for describing structured and semi-structured data. An earlier standard, ASN.1, has been used by the National Center for Biotechnology Information for years to transfer and integrate structured data and has more recently been utilized by data resources such as BIND⁶². Most of the main database providers now make their data (for example, sequences, structures, gene expression profiles and PubMed records) available in XML. Nearly all software vendors are implementing a standard suite of extensions based on XML and web services that make it easy to publish and exchange XML data. This common software base will revolutionize the way data is accessed and used online by liberating data from the software applications that created it^{63,64}.

scores. Bafna *et al.* go on to describe an approach to formalize this and other expert opinions and include them in a complex scoring algorithm. The information required to implement the algorithm is substantial, such as the need to specify the probability of peptide-fragmentation patterns (which are instrument dependent), but such information may be essential to achieve the goal of better operator-independent peptide identifications in complex mixtures.

Another challenge to automating proteomics with complex mixtures is to decide when even the best match of a scoring algorithm is simply not good enough. Better yet would be an approach to state the certainty that an identified match is correct, because not all assignments between peptides and a candidate will be correct. Indeed, with complex mixtures the peptides without correct assignment will likely greatly outnumber those with correct assignments⁴⁷. Establishing a criteria for acceptance overall therefore becomes the main focus of automated proteomics. What criteria should be used to decide whether to accept or reject the assignment deemed 'most-likely' by the scoring algorithm? It is generally assumed that higher-scoring assignments are more likely to be correct than lower-scoring assignments, and so it is common to designate a single score threshold above which all assignments will be accepted. But unlike true *P* values, the score value conveys no information about the actual quality of the match and so it is not possible to directly ascertain the performance characteristics of any specific choice of threshold. For example, the dogma of accepting hypotheses based on *P* values less than 0.05 means, by definition, 5% of all false tests will be misidentified as true. Without such an interpretation of a scoring algorithm, the quality of a match based on automatic scoring cannot be assessed and errors cannot be controlled.

It is essential that some agreed-upon criteria be developed for reporting the quality of any peptide assignment. Any specific threshold could be characterized by its sensitivity (the rate of accepting accurate peptides assignments) and its specificity (the rate of rejecting inaccurate peptide assignment). Of course the specificity and sensitivity of any threshold will depend on the mixture and the

sequence database, because these will affect the distribution of scores among true and false matches. One proposal to control the performance of automated matching has been given by Keller *et al.*⁴⁷ who make the observation that it may be possible to determine the sensitivity and specificity of any assignment using unsupervised, model-based clustering techniques. Keller *et al.* estimate the reference distributions of the correct and incorrect assignments within any experiment. Importantly, their proposal identifies thresholds in an experiment- and database-dependent manner so that a series of experiments can use comparable criteria. In essence, Keller *et al.* describe an approach that may allow a scoring algorithm to be converted into *P* value-like quantities that can then be used to control error rates.

Pattern matching without protein identification

Recently, substantial attention has been given to using chromatography-based proteomics to measure the concentration of low molecular weight peptides in complex mixtures, such as plasma or sera. These technologies commonly use time-of-flight (TOF) spectroscopy with matrix-assisted or surface-enhanced laser desorption/ionization, to produce a spectrum of mass-to-charge (*m/z*) ratios that can be analysed in order to identify unique signatures from its chromatography pattern. Each *m/z* value of the spectrum reflects the abundance of possibly many peptides having a similar mass. Thus, with complex mixtures, these TOF methods are not able to identify individual peptides.

When used with complex mixtures, analysis methods are intended to identify peaks, or features, of the spectrum that can segregate identifiable groups; in this way they are similar to unsupervised learning approaches commonly used when evaluating expression arrays⁴⁸⁻⁵¹. However, because of experimental variation of those spectra, expression array clustering methods are appropriate only if alignment and peak identification and selection algorithms are first used. Adam *et al.*⁴⁸ take this approach to ensure that the features they identify as important are actual peaks. Another approach, used for example by Petricoin *et al.*^{49,50}, is to avoid peak identification all together and accommodate experimental variation in the clustering algorithm. By ignoring peak identification, the resulting classification may produce an algorithm more suitable for prediction, but the features identified may not correspond to actual peaks at all and so this approach may be less useful if eventual peptide identification is the goal.

Even though the TOF algorithms have not yet led to peptide identification, this factor does not greatly limit their utility for identifying newer and far more accurate approaches for medical diagnostics, because diagnosing disease is a problem of prediction rather than of aetiology. Algorithms that have potential clinical relevance have already been identified by Petricoin *et al.*⁵⁰ and Adam *et al.*⁴⁸ for diagnosing ovarian and prostate cancer, respectively. The excitement surrounding these TOF technologies is also due in part to their requiring only very small volumes of specimen (typically less than 50 μ l) to generate their spectra. This is especially true for studies that rely on limited, and therefore precious, supplies of archival specimens⁵². The efficiency of the TOF approaches, and their demonstrated ability to generate highly accurate diagnostic tests in case-control studies, may provide considerable advantages for this technology compared with others for the development of medical diagnostics.

Conclusions and future challenges

Proteomics is a powerful, post-genome paradigm that seeks to describe and explain what Erwin Chargaff called the "immensely diversified phenomenology" of cells and organisms⁵³. Beyond the enumerations and characterizations of different proteomes lies the elucidation of macromolecular interactions, complexes and networks. Informatics will play a crucial role in working towards these goals. Should we be optimistic? To paraphrase

N. K. Hayles, "... annotations, insofar as they represent informational patterns abstracted from their instantiation in a biological substrate, can never fully capture the embodied actuality, unless they are as prolix and noisy as the body itself"⁵⁴. Well, we shall do the best we can. □

doi:10.1038/nature01515

1. Cambridge Healthtech Institute Conference on Human Proteome Project, 2–4 April 2001, McLean, Virginia <<http://www.healthtech.com/2001/hpr/index.htm>> (2001).
2. Krishna, R. G. & Wold, F. Post-translational modification of proteins. *Adv. Enzymol. Relat. Areas Mol. Biol.* **67**, 265–298 (1993).
3. Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nature Rev. Genet.* **2**, 869–878 (2001).
4. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243 (2002).
5. Dayhoff, M. O. & Eck, R. V. MASSPEC: a computer program for complete sequence analysis of large proteins from mass spectrometry data of a single sample. *Comput. Biol. Med.* **1**, 5–28 (1970).
6. Anderson, N. G., Matheson, A. & Anderson, N. L. Back to the future: the human protein index (HPI) and the agenda for post-proteomic biology. *Proteomics* **1**, 3–12 (2001).
7. Boguski, M. S. Bioinformatics. *Curr. Opin. Genet. Dev.* **4**, 383–388 (1994).
8. Boguski, M. S. The turning point in genome research. *Trends Biochem. Sci.* **20**, 295–296 (1995).
9. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965).
10. Dayhoff, M. O. Computer aids to protein sequence determination. *J. Theor. Biol.* **8**, 97–112 (1965).
11. Doolittle, R. F. Some reflections on the early days of sequence searching. *J. Mol. Med.* **75**, 239–241 (1997).
12. Shortliffe, E. *et al.* (eds) *Medical Informatics: Computer Applications in Health Care and Biomedicine* (Springer, New York, 2000).
13. Hieter, P. & Boguski, M. Functional genomics: it's all how you read it. *Science* **278**, 601–602 (1997).
14. Duyk, G. M. Sharper tools and simpler methods. *Nature Genet.* **32**(Chipping Forecast II Suppl.), 465–468 (2002).
15. Kohane, I. S., Kho, A. T. & Butte, A. J. *Microarrays For an Integrative Genomics* (Massachusetts Institute of Technology Press, Cambridge, MA, 2003).
16. Potter, J. D. At the interfaces of epidemiology, genetics and genomics. *Nature Rev. Genet.* **2**, 142–147 (2001).
17. McClatchey, K. D. (ed.) *Clinical Laboratory Medicine* (Lippincott, Philadelphia, 2002).
18. Huang, J. *et al.* Effects of ischemia on gene expression. *J. Surg. Res.* **99**, 222–227 (2001).
19. Craven, R. A. & Banks, R. E. Laser capture microdissection and proteomics: possibilities and limitation. *Proteomics* **1**, 1200–1204 (2001).
20. Craven, R. A. & Banks, R. E. Use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis. *Methods Enzymol.* **356**, 33–49 (2002).
21. Margolin, J. From comparative and functional genomics to practical decisions in the clinic: a view from the trenches. *Genome Res.* **11**, 923–925 (2001).
22. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300 (1995).
23. Dayhoff, M. O. & Eck, R. V. *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, MD, 1966).
24. Smith, T. F. The history of the genetic sequence databases. *Genomics* **6**, 701–707 (1990).
25. Bairoch, A. & Boeckmann, B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19**(Suppl.), 2247–2249 (1991).
26. Maglott, D. R. *et al.* NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28**, 126–128 (2000).
27. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
28. Bader, G. D. *et al.* BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **29**, 242–245 (2001).
29. Adkins, J. N. *et al.* Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteomics* **1**, 947–955 (2002).
30. Kratz, A. & Lewandowski, K. B. Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Normal reference laboratory values. *N. Engl. J. Med.* **339**, 1063–1072 (1998).
31. Jung, E. *et al.* Annotation of glycoproteins in the SWISS-PROT database. *Proteomics* **1**, 262–268 (2001).
32. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
33. Chakravarti, D. N., Chakravarti, B. & Moutsatsos, I. Informatic tools for proteome profiling. *Biotechniques* **32**(Comput. Proteomics Suppl.), S4–S15 (2002).
34. Liebler, D. C. *Introduction to Proteomics* (Humana, Totowa, NJ, 2002).
35. The Association of Biomolecular Resource Facilities. Delta Mass: A Database of Protein Post Translational Modifications <<http://www.abrf.org/index.cfm/dm.home>> (2002).
36. Wilkins, M. R. *et al.* High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.* **289**, 645–657 (1999).
37. Creasy, D. M. & Cottrell, J. S. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2**, 1426–1434 (2002).
38. Choudhary, J. S. *et al.* Matching peptide mass spectra to EST and genomic DNA databases. *Trends Biotechnol.* **19**(Suppl.), S17–S22 (2001).
39. Choudhary, J. S. *et al.* Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1**, 651–667 (2001).
40. Bafna, V. & Edwards, N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **17**(Suppl.) S13–S21 (2001).
41. Eng, J., McCormack, A. & Yates, J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
42. Fenyo, D. Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* **11**, 391–395 (2000).
43. Field, H. I., Fenyo, D. & Beavis, R. C. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2**, 36–47 (2002).
44. Perkins, D. N. *et al.* Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
45. Efron, B. & Tibshirani, R. Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23**, 70–86 (2002).
46. Pepe, M. S. *et al.* Selecting differentially expressed genes from microarray experiments. *Biometrics* (in the press).
47. Keller, A. *et al.* Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
48. Adam, B. L. *et al.* Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62**, 3609–3614 (2002).
49. Petricoin, E. F. III *et al.* Serum proteomic patterns for detection of prostate cancer. *J. Natl Cancer Inst.* **94**, 1576–1578 (2002).
50. Petricoin, E. F. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577 (2002).
51. Qu, Y. *et al.* Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin. Chem.* **48**, 1835–1843 (2002).
52. Pepe, M. S. *et al.* Phases of biomarker development for early detection of cancer. *J. Natl Cancer Inst.* **93**, 1054–1061 (2001).
53. Judson, H. *The Eighth Day of Creation: Makers of the Revolution in Biology* expand. edn (Cold Spring Harbor Laboratory Press, New York, 1996).
54. Hayles, N. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (University of Chicago Press, Chicago, 1999).
55. Bonini, P. *et al.* Errors in laboratory medicine. *Clin. Chem.* **48**, 691–698 (2002).
56. Narayanan, S. The preanalytic phase. An important component of laboratory medicine. *Am. J. Clin. Pathol.* **113**, 429–452 (2000).
57. Spellman, P. T. *et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, 46 (2002).
58. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genet.* **29**, 365–371 (2001).
59. Editorial. Coming to terms with microarrays. *Nature Genet.* **32**, 333–334 (2002).
60. Ball, C. *et al.* Standards for Microarray Data. *Science* **298**, 539 (2002).
61. Orchard, S., Kersey, P., Hermjakob, H. & Apweiler, R. The HUPO proteomics standards initiative meeting: towards common standards for exchanging proteomics data. *Comp. Funct. Genom.* **4**, 16–19 (2003).
62. Bader, G. D. & Hogue, C. W. BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465–477 (2000).
63. Abiteboul, S., Buneman, P. & Suciu, D. *Data on the Web: From Relations to Semistructured Data and XML* (Morgan Kaufmann, San Francisco, 2000).
64. Coyle, F. *XML, Web Services, and the Data Revolution* (Addison-Wesley, Boston, 2002).

Acknowledgements We thank L. Hartwell, J. Potter and G. Omenn for stimulating discussions and J. Gray, J. Pounds and L. Geer for valuable suggestions and critical readings of the manuscript.

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Senior European Sales Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)

Scandinavia: Silke Opstrup (4994)

France/ Belgium:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harakakamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Hideki Watanabe

e-mail: h.watanabe@naturejp.com

naturejobs

Eastern promises

It's a simple case of supply and demand. Denmark has had difficulty recruiting researchers into the country, whereas China has an abundance of skilled postgraduates. Novo Nordisk, the Copenhagen-based drug company, has taken measures to balance this discrepancy by opening a research division in Beijing.

Many Western recruiters have long recognized that there is a pool of potential talent in China. But historically they have had difficulty evaluating Chinese scientists who haven't studied at Western universities or published in English-language journals.

Peter Kurtzhals, senior vice-president of research at Novo Nordisk, solved this problem by bringing in someone he describes as "truly bicultural" — Bao Ping Wang, a Chinese scientist who spent 14 years in the United States, most of them at Harvard. That appointment was crucial to making Novo's Chinese office work, says Kurtzhals. "Wang understands Western culture and has his network back in China and can evaluate local talent." As a result, the office has already grown to about 20 research scientists.

Kurtzhals notes that at the moment growth isn't an issue. Novo, like most drug companies, is biding its time, waiting for the global economy to improve before it undergoes any massive recruitment initiative. But establishing an outpost in China now could offer dividends in the future. When the company feels like growing again, it will already have a system in place to evaluate China's scientific talent. And its Beijing base may provide a way to expand its market into China.

Given those opportunities, it won't be surprising to hear about other companies establishing similar outposts. Nor will it be a shock to hear Western scientists worrying about jobs in their home country being sent east.

Paul Smaglik

Naturejobs editor



Contents

REGIONS

The lure of San Diego p240

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

California dreaming San Diego

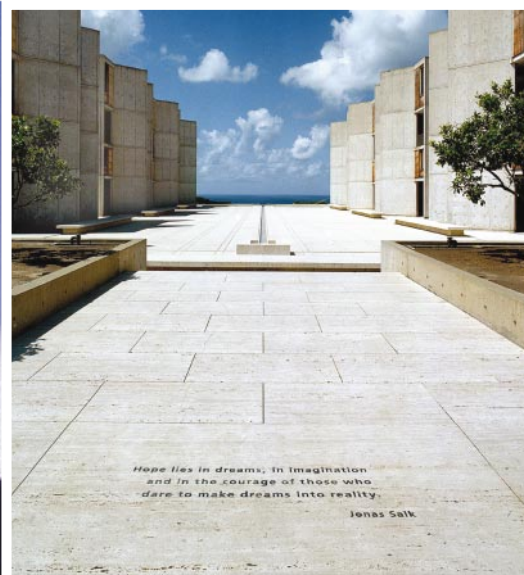
Long-term view: people may come to the Salk Institute (right) and its neighbours for research reasons, but the climate and the environment make it hard to leave.

“I asked Jonas Salk why he came here and he said the weather,” says Joe Panetta, president and chief executive of Biocom, a biotech trade association in San Diego, California. “I asked Francis Crick why he came here and he said the weather.” Since those two life-science luminaries made the move, thousands have followed. Starting in the 1960s, the flat-topped hill of Torrey Pine mesa and other environs near San Diego and La Jolla have become a centre of life-science research.

The mesa alone hosts the Salk Institute, the Scripps Research Institute, the Burnham Institute and the life-science campus of the University of California, San Diego (UCSD), among others. Nestled in the hills and valleys nearby are hundreds of biotech companies, a growing life-science tools presence, and signs of expansion by pharmaceutical companies.

But despite the sun that shines all but ten or so days a year, there are clouds on the employment horizon. A stagnant US stock market may affect investment in companies that have not brought a product to market and reduce contributions to the non-profit research institutions' endowments. A state-government deficit means that the politicians in Sacramento may be unable to prime the economic pump.

Still, there are some reasons for optimism. Marney Cox, chief economist for the San Diego Association of Governments, says that overall employment in the San



Hope lies in dreams, in imagination
and in the courage of those who
dare to make dreams into reality.
Jonas Salk

Diego region has grown consistently during the past 15 years. Although the telecommunications industry has recently shed jobs, software firms have hired more staff and biotech has more or less held its own, with some layoffs being absorbed by other companies.

Cox sees potential in the life sciences because — although only one biotech company, IDEC Pharmaceuticals, is ready to manufacture an approved drug — several others will be wrapping up their clinical trials within the year and approaching US Food and Drug Administration (FDA) approval. He thinks that the city's military history will help to bring in funds for biodefence, which is one of the biggest areas of increase in the 2003 US R&D budget.

ACADEMIC OPTIMISM

Despite the economic problems faced by both state and US federal governments, Lawrence Goldstein, Howard Hughes Medical Institute investigator at the University of California, San Diego, remains bullish about San Diego. He has no regrets that he left Harvard University for UCSD ten years ago. He especially enjoys the concentration of research institutions, all of which host seminars that regularly draw world-class scientists and are open to their neighbouring institutions and companies alike. The proximity — and the sunshine — means he can reach many meetings by taking a short walk. At Harvard, he says, “it was a pain in the neck to get to the next building”, especially during the winter.

He believes that UCSD will eventually double in size, although probably not within the ten-year timescale that its current planners predict. A new school of pharmacy has recently opened. And as far as the average researcher is concerned, Goldstein doesn't think California's deficit will have much impact. “Postdocs don't get state money anyway,” he says.

Jeffrey Kelly, vice-president of academic affairs at the Scripps Research Institute, is, if anything, even less concerned about money. The Scripps doesn't depend on state finances, and all of its researchers compete well for federal dollars. “If they are in the top 15% of science, they'll be fine,” Kelly says. “If not, you shouldn't hire them.”

Lawrence Goldstein doesn't regret his move from Harvard, and doubts whether California's financial problems will affect the average researcher.

Extension classes

The University of California, San Diego, could be seen as biotech's classroom. Its bioscience division offers courses that educate scientists about every stage of the drug-development pipeline, from bioinformatics to regulatory affairs. About

half of the students have advanced degrees and about 80% work in the local biotech industry. The idea is catching on. One course in computer-aided drug design drew people from Sweden, Quebec, Taiwan and India.

P.S.

♦ www.extension.ucsd.edu



There always seems to be room for qualified people to join the 280 faculty members, 800 postdocs and 170 graduate students, Kelly notes: "We always have opportunities." He prefers to find the best available people and then build up programmes around them, rather than identify programmes to recruit scientists into targeted areas. Even graduate students get that treatment. Last year he brought in a promising graduate student who wanted to study Gaucher's disease, and promptly recruited a world expert on this genetic blood disorder. By the end of the year, the student had a paper published in *Proceedings of the National Academy of Sciences*.

Even though he doesn't like to target specific areas, Kelly would like to add more neuroscience and build up translational medicine — ideally by bringing in MD/PhDs who have spent time in both academia and the pharmaceutical industry.

Unlike its neighbour the Scripps, the Salk does not expect to expand much during the next few years. But Polly Murphy, vice-president of intellectual property and technology transfer, expects retirements to create about 12 or 14 vacancies, and the institution's pool of about 300 postdocs is constantly replenished.

Nancy Beddingfield, director of institutional relations at the Burnham Institute in La Jolla, says that space is a limitation. "There's no more land here on the mesa," she says. Apart from UCSD, institutions there are not allowed to build higher than 10.6 metres. They are also limited by the need to preserve the country's largest collection of Torrey pines — the squat, gnarled evergreens that give their name to the mesa. "We built around the pine trees," says Beddingfield.

INDUSTRY ANTICIPATION

Another force making space scarce and expensive on a nearby mesa is the advance of the pharmaceutical industry. Scott Struthers, director of exploratory research at Neurocrine, can look out of his office window and watch the progress of one of the eight or more Pfizer buildings under construction.

Although not welcoming the inevitable rent increase, Struthers is one of several people who believe a bigger pharmaceutical presence will give the area more stability. If a researcher has bad luck with small



Staying in touch: researchers at Invitrogen's Carlsbad buildings have strong links with the University of California, San Diego.

companies that run out of cash or fail to get products to market, they may be able to find work at Pfizer or at Merck, which is also expanding into the area.

Struthers suspects that Neurocrine employees won't be among them. The company has an anti-insomnia drug in the final stage of trials and, if it wins FDA approval, Pfizer will help to market it. That success will mean more clinical-development people as well as basic researchers to fill the pipeline. "We're planning on doubling research in the next 3–5 years," he says. Struthers, like his academic colleagues, enjoys life on the mesas. "We walk to the seminars," he says. "A lot of us teach."

Some companies are locating farther away, partly to avoid La Jolla's land prices. Research-tools company Invitrogen, for example, built a 300,000-square-foot production facility in Carlsbad, where it has R&D facilities. Still, employees manage to interact with the region's key players. John Carrino, vice-president for R&D, says that the company uses researchers at UCSD to help evaluate technologies that it is developing or considering licensing. The university also provides peer review for its external grant programme (see 'Tool grants', left).

Structural GenomiX, in San Diego, also has a lot of interaction with the academic research community, says Stephen Burley, its chief scientific officer and senior vice-president of research. Nobel laureate Sydney Brenner, for example, is an adviser to the company, which is part of the US structural genomics initiative, supervised by the National Institutes of Health. Some Structural GenomiX researchers work in labs selected by their non-profit partners, and are paid with federal dollars. Others work on proteins that the company has selected as possible therapeutic targets.

The boundaries between academic and industrial research in San Diego have become "increasingly blurred", says Bruce Stevenson, vice-president of academic affairs at the Salk. He notes that the mix of biotech, pharmaceuticals and academia gives graduate students more career paths to choose from in the area.

If that is not incentive enough, there is always the weather and the ocean views. "Sitting and looking at sunsets over the Pacific ain't bad," Stevenson says.

Paul Smaglik is editor of *Naturejobs*.

Bruce Stevenson finds the boundaries between academic and industrial research are blurred.



Tool grants

Invitrogen, the research-tools company in Carlsbad, California, has a relatively new external granting programme that, its director says, appeals especially to younger scientists. Most academic scientists who are industry-minded shoot first for therapeutics or big academic grants. But younger scientists who are adept at developing

reagents may consider applying for the company's research-tools grant, as a form of early funding. Every year Invitrogen awards about \$100,000 each to scientists proposing the creation of research tools. Last year, 150 people made pre-proposals, 100 were asked to submit full ones, and about 20 were funded. P.S. ♦ www.invitrogen.com